Study of the dynamics of gene expression by mathematical modelling and systems approaches

Submitted for the degree of Doctor of Philosophy

Thomas Brown
Trinity College, University of Oxford

Hilary Term, 2019

Abstract

Transcription by RNA Polymerase II (RNA PolII) is a fundamental process at the heart of gene expression, defining the phenotypes of cells. Transcription of genes does not exist in isolation, but occurs through a complex environment of DNA wrapped around chromatin, with many factors affecting the surrounding environment and association with the polymerase. The process of transcription is more complex than perhaps originally thought, with a number of factors affecting individual steps of the process of reading the DNA template. Rather than thinking of a gene as simply being on or off, the expression of a gene is regulated and fine-tuned by the modifications to the chromatin landscape, cross-talk between the chromatin and the elongation complex, pervasive non-coding transcription and the association of factors with the polymerase. By developing mathematical models of the transcription and RNA processing networks, we determine the effects of these factors on establishing the dynamics of native gene expression. Constructing best-fit models and identifying key parameters which dictate the changes to quantitative data upon genetic intervention, we identify a feedback mechanism to accommodate for changes in the rates of transcript production and polymerase stalling as a predictor of the processivity of individual genes. A number of factors are shown to control the speed and processivity of polymerase in *Saccharomyces cerevisiae* and HeLa cells, establishing a network of co-operative and antagonistic interactions with the RNA PolII complex. This work suggests that the dynamics of transcription are gene-specific, with the chromatin landscape and the polymerase-associating factors unique to each locus in the mammalian genome.

Acknowledgements

A number of people have contributed to this thesis, without whom this work would be a skeleton of its current incarnation. Many thanks to Andrew and Jane for guiding and encouraging me through the last 4 years, it has been great fun being part of their groups. Thank you to everyone in the Mellor and Angel labs for creating a fantastic environment in which to work. In particular, thanks to Harry, Ronja, Françoise, Ülkü, Meredith and Shidong whose data has meant that there is some content to this thesis. To Charlie, who added much more rigour to the bioinformatics and brought some excellent donuts to the lab. To Silvia and Ülkü for bringing out my inner Regina George. Who knew how important it is to have a girls' night every now and then to let off some steam? Harry, for ensuring the cakes in the lab were always up to standard. Last, and especially least, to Struan. Very few things that I brought into the lab are not somewhere under that pile of "novel ideas" and "bioinformatic brainwaves" on your desk, but I can forgive the loss of a few pens. I'm sure we're all looking forward to reading 'Orphans of the Tide' when it's 1p on Kindle.

Special thanks to Hannah for allowing me to continually expand my ever-growing list of hobbies and encouraging me towards a life of professional procrastinator.

Contents

1 Introduction **1**
1.1 Transcription and gene expression . . . 1
1.1.1 Systems approaches to studying gene expression . . . 2
1.1.2 Outline . . . 2
1.2 Types of transcription units . . . 3
1.2.1 Protein coding genes . . . 3
1.2.2 Non-coding RNA in *S. cerevisiae* . . . 4
1.3 RNA polymerase II transcription . . . 7
1.3.1 Transcription initiation . . . 7
1.3.2 Promoter escape . . . 10
1.3.3 Transcription elongation . . . 11
1.3.4 Transcription termination . . . 12
1.4 RNA processing . . . 13
1.4.1 RNA export . . . 13
1.4.2 RNA degradation . . . 13
1.5 The dynamic nature of transcription . . . 15
1.5.1 The Yeast Metabolic Cycle . . . 16
1.5.2 Gene expression under different carbon sources . . . 17
1.5.3 The *GAL* locus . . . 17
1.6 Techniques used to study transcription dynamics . . . 19
1.6.1 Single-gene analysis . . . 20
1.6.2 Genome-wide analysis . . . 22
1.7 Factors affecting transcription and gene regulation . . . 26
1.7.1 Transcription and elongation factors . . . 27
1.7.2 DNA and chromatin . . . 27
1.7.3 Non-coding transcription . . . 28
1.7.4 RNA processing factors . . . 29
1.8 Distributions of quantitative data reveal more information than summary statistics . . . 29
1.8.1 Constructing models to obtain distributions . . . 30
1.8.2 Existing models of transcription . . . 31
1.9 Data included and analysed in this study . . . 33
1.10 Aims of research . . . 35

2 Gene regulation is dynamic upon change of media **38**
2.1 Overview . . . 38
2.2 Deep sequencing techniques allow for analysis of transcript dynamics . . . 39
2.2.1 The Yeast Metabolic Cycle demonstrates cycling sub-classes of genes 40
2.2.2 MNase-digestion followed by deep sequencing shows locations of nucleosomes . . . 41

2.2.3 Nascent and mature RNA sequencing give high-resolution views of gene expression dynamics . . . 43
2.3 Plus one nucleosome position moves with gene expression activity . . . 44
2.3.1 Nucleosomes move into the gene body at activated genes . . . 47
2.3.2 Nucleosomes move towards the promoter at repressed genes . . . 47
2.4 Discussion . . . 47

3 RNA production dynamics are modulated through histone acetylation 54
3.1 Transcription occurs through a complex environment . . . 54
3.1.1 Transcription units are associated with unique transcription environments . . . 56
3.2 Modelling RNA FISH data reveals distinct transcription dynamics . . . 57
3.2.1 RNA FISH . . . 57
3.2.2 Promoter dynamics . . . 58
3.2.3 Polymerase elongation . . . 59
3.2.4 Degradation of RNA . . . 60
3.2.5 A model of RNA production . . . 60
3.2.6 Gillespie simulation of RNA distributions . . . 63
3.2.7 Parameter estimation via Approximate Bayesian Computation . . . 64
3.2.8 Determining degradation rates . . . 65
3.3 Transcription of *GAL1-10* locus in glucose and galactose media . . . 66
3.4 Does antisense transcription affect sense transcription dynamics? . . . 67
3.5 Altering acetylation levels mimics effects of altering antisense transcription levels . . . 69
3.6 Do histone methyltransferases have roles in governing transcription dynamics? . . . 72
3.6.1 Genes exhibit a range of transcription dynamics . . . 72
3.6.2 H3K4me3 does not control transcription dynamics . . . 75
3.7 Conclusions . . . 78
3.7.1 Antisense and histone acetylation affect rates of transcription . . . 78
3.7.2 Transcription dynamics are distinct at individual genes . . . 80

4 Polymerase activity is distinct across the genome 81
4.1 Introduction . . . 81
4.1.1 From single genes to genome-wide analysis . . . 81
4.1.2 High resolution views of polymerase activity . . . 82
4.1.3 The dynamics of RNA polymerase II . . . 83
4.1.4 Aims . . . 86
4.2 Polymerase profiles are distinct across the genome . . . 87
4.3 Polymerase profiles are independent of expression levels . . . 89
4.4 Inhibiting elongation factors alter polymerase dynamics . . . 93
4.4.1 Spt4 assists polymerase through early transcription . . . 93
4.4.2 Backtrack release determines polymerase location . . . 95
4.4.3 What role does Paf1 have in regulating early transcription? . . . 95
4.5 How do histone writers influence distributions of polymerase? . . . 96
4.5.1 The role of histone acetylation on transcription dynamics . . . 96

4.5.2 Histone methyltransferases do not have pronounced effects on sense transcription . . . 98
4.6 Polymerase profiles in mammalian cells . . . 100
4.6.1 Post-translational modifications of elongation factors determine polymerase profiles . . . 103
4.7 Discussion . . . 104

5 RNA polymerase II exists in a number of states 108
5.1 A stochastic model of polymerase elongation . . . 108
5.1.1 Aims . . . 109
5.2 Models of polymerase elongation . . . 110
5.2.1 Rules of the models of transcription . . . 111
5.2.2 A list of analysed models . . . 115
5.2.3 Calculating the observed elongation rate . . . 117
5.3 Fitting simulated distributions to NET-seq . . . 121
5.4 Backtracking and two windows are components of the best-fit model . . . 123
5.4.1 Relating parameters to polymerase distributions . . . 125
5.4.2 Sub-groups of genes are best-fit by different models of transcription 130
5.5 Transcription dynamics dictate polymerase distributions . . . 131
5.5.1 Altered distributions reflect altered processivity . . . 135
5.6 Transcription dynamics are distinct across the genome . . . 137
5.7 Using modelling to obtain numerical transcription rates . . . 141
5.7.1 Spt4 assists polymerase processivity . . . 142
5.7.2 Dst1 determines polymerase dynamics in the proximal window . . . 144
5.7.3 Paf1 is a pause-inducing factor proximal to the promoter . . . 145
5.7.4 Xrn1 inhibits transcription processivity . . . 147
5.7.5 Tyrosine 1 of the RNA PolII CTD has a role in promoter-proximal dynamics . . . 149
5.7.6 The effects of histone modifications on transcription dynamics . . . 151
5.8 Transcription dynamics in HeLa cells . . . 159
5.8.1 Transcription dynamics are conserved between *S. cerevisiae* and HeLa cells . . . 159
5.8.2 Transcription dynamics are distinct across the human genome . . . 162
5.8.3 CDK9 controls mobility of promoter-proximal polymerase . . . 165
5.9 Discussion . . . 167
5.9.1 Transcription processivity is determined by many factors . . . 171

6 Conclusions and future work 172
6.1 Summary of results . . . 172
6.1.1 Nucleosomes are highly dynamic upon changes of gene expression . 172
6.1.2 Chromatin acts to modulate transcription dynamics . . . 173
6.1.3 Transcription dynamics are distinct across eukaryotic genomes . . . 174
6.1.4 Many factors work cooperatively and antagonistically to determine the parameters of transcription . . . 176
6.1.5 Histone modifications modulate polymerase elongation . . . 177
6.1.6 RNA PolII transcribes in the face of multiple obstacles . . . 178
6.2 Future work . . . 179

7 Appendix **181**
7.1 Nucleosome locations of YMC classes of genes upon galactose shift 181
7.2 Derivation of the probability generating function for bursting protein production 185
7.3 Algorithm for simulation of RNA FISH data 189
7.4 Algorithm for simulation of NET-seq distributions 191
7.5 Results from simulations of mNET-seq data after treatment with DRB . . 194

List of Figures

1.1 A snapshot of nascent transcription . . . 6
1.2 Outline of the stages of transcription . . . 7
1.3 The CTD code . . . 10
1.4 Pathways of RNA processing . . . 14
1.5 The network of galactose metabolism . . . 18
1.6 The *GAL* locus . . . 19
1.7 Schematic of polymerase sequencing methods . . . 23

2.1 The yeast metabolic cycle . . . 41
2.2 Schematic of MNase-seq . . . 42
2.3 All genes nucleosome position changes . . . 46
2.4 Cluster 1 nucleosome position changes . . . 48
2.5 Cluster 2 nucleosome position changes . . . 49
2.6 Cluster 3 nucleosome position changes . . . 50
2.7 Cluster 4 nucleosome position changes . . . 51

3.1 Types of transcription units . . . 55
3.2 A schematic of RNA FISH . . . 58
3.3 Bursting kinetics at *GAL10* . . . 59
3.4 A model of RNA production . . . 62
3.5 Schematic of the inference of parameters from FISH modelling . . . 65
3.6 Layout of *GAL1* locus with *GAL1* antisense mutation . . . 67
3.7 *GAL1* sense and antisense levels . . . 67
3.8 FISH data from engineered strains . . . 69
3.9 Non-bursting models give poor fits to data . . . 70
3.10 Inferred sense transcription rates across a locus with high and low levels of antisense transcription . . . 70
3.11 Decay rates of *GAL1* mRNA . . . 71
3.12 Inferred transcription rates in the absence of histone deacetylase Set3 . . . 71
3.13 Endogenous genes FISH data and fits I . . . 73
3.14 Endogenous genes FISH data and fits II . . . 74
3.15 Endogenous genes FISH data and fits III . . . 75
3.16 Endogenous genes FISH data and fits IV . . . 76
3.17 Transcription rates for endogenous genes . . . 77
3.18 The Set1 complex . . . 77
3.19 Transcription rates in WT and spp1 strains . . . 78

4.1 Schematic of NET-seq protocol . . . 83
4.2 A schematic of promoter-proximal pausing . . . 86
4.3 NET-seq profiles in *S. cerevisiae* . . . 87
4.4 NET-seq profiles in *S. cerevisiae* from Stirling Churchman . . . 89
4.5 Clustering of NET-seq data through the YMC . . . 91

4.6 Comparison of NET-seq profiles with RNA-seq levels . . . 92
4.7 NET-seq profiles in *S. cerevisiae spt4*Δ . . . 93
4.8 PRO-seq profiles in *S. cerevisiae spt4*Δ . . . 94
4.9 NET-seq profiles in *S. cerevisiae dst1*Δ . . . 95
4.10 NET-seq profiles in *S. cerevisiae paf1*Δ . . . 96
4.11 NET-seq profiles in *S. cerevisiae eaf3*Δ . . . 97
4.12 NET-seq profiles in *S. cerevisiae rco1*Δ . . . 98
4.13 NET-seq profiles in *S. cerevisiae set1*Δ . . . 99
4.14 NET-seq profiles in *S. cerevisiae spp1*Δ . . . 99
4.15 NET-seq profiles in *S. cerevisiae set2*Δ . . . 100
4.16 mNET-seq profiles in HeLa cells . . . 102
4.17 mNET-seq profiles compared to expression . . . 103
4.18 mNET-seq profiles following inhibition of CDK9 . . . 105

5.1 States of polymerase model . . . 114
5.2 Akaike Information Criteria for transcription models . . . 123
5.3 Transcription dynamics determining polymerase levels . . . 126
5.4 Comparison of inferred parameters from FISH with NET-seq polymerase levels . . . 127
5.5 Transcription dynamics determining polymerase proximal-distal ratio . . . 129
5.6 Transcription dynamics determining polymerase processivity . . . 129
5.7 Akaike Information Criteria following profile clustering . . . 131
5.8 Simulated NET-seq distributions under altered dynamics . . . 132
5.9 Proximal-distal ratio of polymerase after altering initiation and elongation 133
5.10 Simulated NET-seq distributions under altered stalling dynamics . . . 134
5.11 Proximal-distal ratio of polymerase after altering stalling and backtracking 136
5.12 Simulated NET-seq distributions after altered backtracking . . . 137
5.13 Difference in processivity for altered dynamic rates . . . 138
5.14 Simulation parameters from clustered NET-seq distributions 1 . . . 140
5.15 Simulation parameters from clustered NET-seq distributions 2 . . . 141
5.16 Inferred transcription parameters in *spt4*Δ cells . . . 143
5.17 Inferred transcription parameters in *dst1*Δ cells . . . 146
5.18 Inferred transcription parameters in *paf1*Δ cells . . . 147
5.19 Inferred transcription parameters in *xrn1*Δ cells . . . 148
5.20 Inferred transcription parameters in Y_1 CTD mutant cells . . . 150
5.21 Inferred transcription parameters in *eaf3*Δ cells . . . 153
5.22 Inferred transcription parameters in *rco1*Δ cells . . . 154
5.23 Inferred transcription parameters in *set1*Δ cells . . . 156
5.24 Inferred transcription parameters in *spp1*Δ cells . . . 157
5.25 Inferred transcription parameters in *set2*Δ cells . . . 158
5.26 Akaike Information Criteria for transcription models of HeLa mNET-seq data . . . 160
5.27 Akaike Information Criteria for clustered HeLa mNET-seq data . . . 161
5.28 HeLa inferred initiation and elongation parameters . . . 163
5.29 HeLa inferred window sizes . . . 164
5.30 HeLa inferred processivity . . . 165
5.31 Raw mNET-seq data and fitted distributions for wild-type and CDK9-inhibited HeLa cells . . . 166
5.32 Parameter differences in HeLa cells after inhibition of CDK9 . . . 167
5.33 A summary of the effect of elongation factors on polymerase dynamics . . . 169

5.34 A summary of the effect of histone modifications on polymerase dynamics . 170

7.1 Ox genes nucleosome position changes 182
7.2 RB genes nucleosome position changes 183
7.3 RC genes nucleosome position changes 184
7.4 Parameter differences in HeLa cells after 5 minutes inhibition of CDK9 . . 194
7.5 Parameter differences in HeLa cells after 5 minutes inhibition of CDK9 . . 194

Chapter 1

Introduction

1.1 Transcription and gene expression

The conversion of DNA into RNA is fundamental step central to the cycle of converting genetic material into functional molecules. Transcription is the process of converting DNA to either a functional messenger RNA, which is then further converted to protein, or a non-coding RNA, which may have further roles as an RNA molecule or be considered spurious. As a fundamental step underpinning both the phenotypes of cells and governing a cell's ability to respond to external cues, gene expression is a highly regulated step, with multiple proteins and networks governing the expression of different genes. The breakdown of this tightly regulated step can result in a number of diseases, such as in cancer development, where the regulation of transcription is known to be disrupted (Cox & Goding 1991, Anderson *et al.* 2002).

Central to transcription is the multi-protein enzyme RNA Polymerase II (RNA PolII), which transcribes all protein-coding genes and a number of non-coding RNA in *Saccharomyces cerevisiae* and other eukaryotic organisms. Many proteins interact with RNA PolII, assisting the passage from start to the end of genes. Studying the dynamics of the steps involved in RNA production assist us in understanding the rate-limiting factors involved in gene expression and how regulatory factors contribute to these dynamics and ultimately gene expression. Whether dynamics are homogeneous between genes, or even across a single gene, remains unknown and the rates of dynamics underpin a fundamental understanding of gene expression.

1.1.1 Systems approaches to studying gene expression

As more quantitative data is produced to study transcription and gene expression, more quantitative tools need to be developed in order to correctly interpret results. Bioinformatic tools comparing genes within a single dataset and across multiple datasets allow us to answer much more detailed questions about the roles of individual factors in a biological setting. We can separate factors which have the same effect at every gene, to those which have different roles at sub-groups of genes or in different networks and this allows us to build a much more accurate picture of the complex network of interactions which exist inside the cell.

Mathematical modelling allows us to infer more detailed results from quantitative data, as we are able to move beyond speculation as to how individual steps of a network affect the final outcome, but can determine this in a quantitative manner. By building models of biological systems based on known interactions from the literature, we can first determine physiological dynamic regimes for each process in the cell and furthermore determine the impact of regulatory factors on all of these dynamics. Using quantitative experimental tools, bioinformatic techniques and mathematical modelling of biological networks together, we are able to make much more definitive statements about the roles of regulatory factors in the cell and how all of these factors work together to set up the fundamental process of transcription, which lies at the heart of biology.

1.1.2 Outline

In this chapter we outline a number of factors that contribute to the complex network which regulates transcription and gene expression. First, we present the range of genes that exist in eukaryotic organisms and how they fit together to regulate each other and control the balance of RNA and protein in the cell. Then we establish the key steps in transcription, the process carried out by the RNA PolII complex from initiation to termination, followed by the processing of the produced RNA molecule. This sets out the known and speculated regulatory steps involving the polymerase complex and how other factors impact this complex during this multi-step process. We then present a broader overview of gene expression in dynamic systems in *S. cerevisiae* and the techniques used

to investigate the dynamics of transcription, both through static steady-state measurements and through varying conditions. We then outline the known mechanisms by which transcription and gene expression are controlled by the factors outlined above and further speculate what further regulatory mechanisms may exist to fine-tune the process of transcription. Finally we outline the importance of using quantitative data to its full capacity and how mathematical modelling serves to act as the perfect tool for this purpose. By moving beyond a binary on/off view of gene expression to a more nuanced idea of the distribution of molecules across genes, genomes and populations of cells, we are able to answer more in-depth questions about how individual elements affect gene expression at much finer detail.

1.2 Types of transcription units

A vast array of types of genes are known to exist in eukaryotic organisms, varying from fully functional to spurious. The crowded nature of genomes such as *S. cerevisiae*'s means that many transcription units form overlapping regions, creating potential regulatory roles for the acts of transcription and the resulting transcripts (Mellor *et al.* 2016). Below are descriptions of a few of the types of transcription units that exist in *S. cerevisisae*, how they are currently understood to be regulated and the reported methods by which they regulate each other.

1.2.1 Protein coding genes

The textbook view of transcription is to convert the genetic templates found in the DNA sequence to complementary messenger RNA (mRNA), which is then translated into proteins which perform functional tasks within the cell. For many transcription units, this is the case, where RNA PolII targets the protein coding regions of the genome and transcribes the sequence which will eventually encode the proteins. This mRNA also contains an untranslated region upstream of the coding template (5' UTR), which is capped with a methylated guanine residue, protecting the 5' end of the mRNA from degradation. The 3' end of the mRNA further contains an untranslated region (3' UTR), which may be cleaved, then the poly-adenylation machinery adds a poly(A) tail to the mRNA, protecting the 3'

end of the mRNA from degradation.

In higher eukaryotes, the majority of genes are made up of functional exon sequences, interspaced with non-coding intronic sequences. These introns must be removed from the mRNA before they can be translated, which is performed by the splicing machinery.

1.2.2 Non-coding RNA in *S. cerevisiae*

A number of annotated stable non-coding transcripts exist in *S. cerevisiae*, which are transcribed at reasonably high levels under normal growth conditions (Xu *et al.* 2009, Pelechano *et al.* 2014). One of the best annotated non-coding RNAs (ncRNAs) in yeast is the *GAL10* long non-coding RNA (lncRNA), which originates in the body of *GAL10* and reads through the shared promoter of *GAL10* and *GAL1*, well into the body of *GAL1* in a sense orientation (Houseley *et al.* 2008, Zacharioudakis & Tzamarias 2017, Lenstra *et al.* 2015). This particular lncRNA is expressed in glucose- and raffinose-containing media and when cells are shifted to galactose-containing media, expression of this lncRNA is switched off as the galactose-metabolising genes are switched on (Lenstra *et al.* 2015), suggesting that the transcription of this ncRNA could be functional in the control of a nearby gene's expression.

Cryptic unstable transcripts

A number of unstable transcription units were identified in *S. cerevisiae* after the depletion of the nuclear RNA degradation machinery. This highlighted hundreds of transcripts which undergo rapid degradation in the nucleus and are not detected by sequencing methods such as RNA-seq at low read-depth (Wyers *et al.* 2005). These transcription units were termed cryptic unstable transcripts (CUTs) and the full list of CUTs has been vastly built on by the Steinmetz lab (Xu *et al.* 2009, Neil *et al.* 2009). A global role for these CUTs has not yet been determined, with a number of individual cases studied for their roles in activating or repressing the expression of surrounding genes (e.g. Kim *et al.* 2012).

Other unstable transcripts

A number of new classes of unstable transcripts have been introduced to the growing list of transcription units. Similar in their discovery to Rrp6-mediated CUTs, Xrn1-sensitive unstable transcripts (XUTs) have also been describe in yeast (van Dijk *et al.* 2011). Xrn1 is part of the cytoplasmic degradation machinery, degrading RNA in a 5' - 3' direction.

Further non-coding transcription units have also been discovered by disrupting the action of the termination factor Nrd1, giving rise to a set of Nrd1-unterminated transcripts (NUTs). These arise through incorrect termination of transcription units mostly from bi-directional transcription at promoters (Schulz *et al.* 2013). Finally, a set of Ssu72-restricted transcripts (SRTs) were described through mutation of the RNA processing factor Ssu72, resulting in the discovery of a number of transcripts also arising from bi-directional promoters (Tan-Wong *et al.* 2012).

These groups of transcription units all contribute to the rich tapestry of genes which cover the majority, if not all, of the yeast genome (Mellor *et al.* 2016).

Antisense transcription

In *S. cerevisiae*, at least 36% of coding transcription units display non-coding transcription originating from the terminator region and read in the reverse orientation to the sense, protein-coding direction (Murray *et al.* 2012, Figure 1.1). Below, we refer to this form of transcription as antisense transcription, where polymerase may read through the gene body and into the promoter in the reverse, or antisense, orientation. This form of transcription has been shown to have diverse and contradictory roles in gene expression, with reports of repressive activity on the corresponding sense gene in *S. pombe* (Eser *et al.* 2016) and modifications to the chromatin state in *S. cerevisiae* and human cells (Murray *et al.* 2015, Mayer *et al.* 2015, Brown *et al.* 2018). However there appears to be no global correlation between the levels of sense transcription and antisense transcription at the same locus (Murray *et al.* 2012, Murray *et al.* 2015).

Recent work has shown that the sense and antisense transcription are performed by different polymerase complexes, which dictates the fate of the produced RNA. Unlike when transcribing a protein-coding gene, the RNA PolII CTD does not exhibit increased

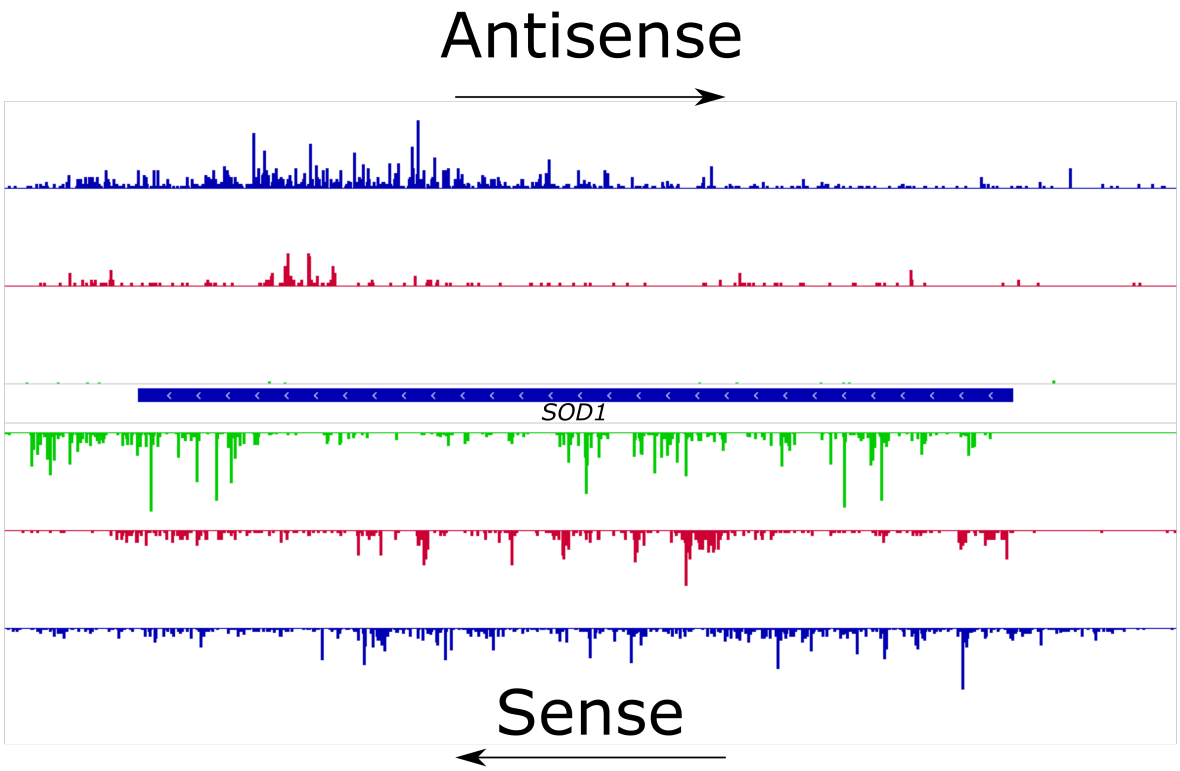


Figure 1.1: Sense and antisense transcription at the gene *SOD1*. Shown are the locations of polymerase actively transcribing on the Watson (above) and Crick (below) strands as measured by NET-seq (blue) and PRO-seq (red). Shown in green is the mature RNA produced from the sense transcription of *SOD1* (below) with the antisense transcript being rapidly degraded (above).

phosporylation of the Serine_2 residue (Murray *et al.* 2015) and the polymerase is shown to associate with different elongation factors (Fischl *et al.* 2017). The Paf1C complex was shown to preferentially bind to transcription of coding regions, assisting with the rapid export of the produced RNA into the cytoplasm for translation. Transcription units depleted for Paf1C were then targeted for nuclear degradation as a mechanism to distinguish coding from non-coding transcripts. This firstly highlights that the textbook idea of transcription by RNA PolII as identified at some genes does not apply to all transcription units and furthermore allows us to investigate whether other factors or modifications associated with the polymerase complex are specific to sub-groups of genes. In this study we will investigate what roles these complexes of different proteins associated with RNA PolII have on the dynamics of the transcription process.

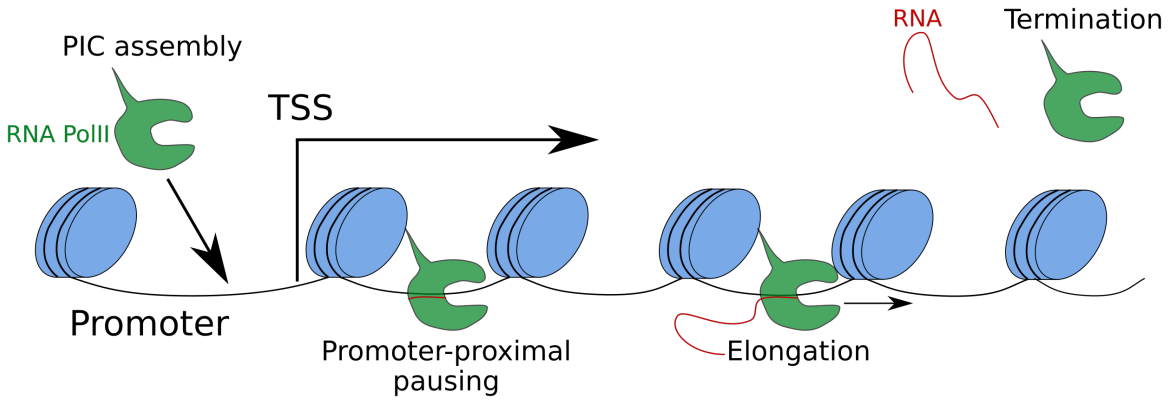


Figure 1.2: The key stages of the transcription process. The Pre-Initiation Complex (PIC) assembles at gene promoters, upstream of the Transcription Start Site (TSS). Transcription begins with a small number of bases being read, before the polymerase may pause in the promoter-proximal region. RNA PolII is then released into processive elongation until it reaches the end of the gene, whereby it undergoes termination and the fully-formed RNA molecule is cleaved.

1.3 RNA polymerase II transcription

The steps taken by polymerase to encode functional RNA from the genetic template in the DNA can be broadly broken into the three steps of initiation, elongation and termination (Figure 1.2), however each of these steps consists of a number of processes, with a number of overlapping transitional stages.

1.3.1 Transcription initiation

In order to begin the process of producing RNA, the pre-initiation complex (PIC) must form at the promoters of genes. The nature by which promoters are marked and identified has been thoroughly studied, but still remains elusive (Venters *et al.* 2011, Ferrari & Strubin 2015). Promoters are marked by the chromatin landscape, with transcription start-sites (TSSs) found abutting nucleosome-depleted regions (NDRs), where the two nucleosomes surrounding the TSS are spaced further apart than the typical ∼150bp repeating pattern of nucleosomes in gene bodies (Fan *et al.* 2010). These NDRs are not exclusive to the promoter regions and are also found in the terminator regions of genes.

It has been suggested that at some genes, the promoter region is occupied by a "fragile" nucleosome, which facilitates the recruitment of transcription factors (Knight *et al.* 2014). These fragile nucleosomes have been particularly associated with the highly-expressed ribosomal-protein genes, whose expression is governed by the recruitment of transcription

factors Rap1, Ifh1, Fhl1 and Hmo1 (Shore & Nasmyth 1987, Kubik *et al.* 2015). Under repressive conditions of these genes, it appears the fragile nucleosomes sitting in the promoter regions of the RPGs is removed and with it the corresponding transcription factors responsible for expression (Rap1, Ifh1, Fhl1) leave the promoters regions (Kubik *et al.* 2015). This gives an example of seemingly more open chromatin, where the NDR is less occupied by nucleosomes is associated with transcriptional repression, with the locations of nucleosomes directly impacting the activity at genes or vice-versa.

The promoters of active genes have furthermore been shown to be highly methylated at the 4th amino acid, a lysine, on histone 3 (H3K4me) (Santos-Rosa *et al.* 2002), however it is not yet clear whether the tri-methylation of H3K4 marks genes to be transcribed or vice-versa (Howe *et al.* 2017). Furthermore, the NDR found at promoters is not always uni-directional, with a number of coding genes sharing a promoter in yeast (e.g. the *GAL1-10* locus, St John & Davis 1981) or causing bi-directional non-coding transcription, termed PROMPTs in mammalian cells (Preker *et al.* 2008). The directionality of transcription at promoters is thought to have evolved, with promoters originally having no directionality, evolving a more favoured direction over time (Jin *et al.* 2017).

Once a promoter or transcription initiation site is selected, the polymerase must bind the promoter as part of the PIC. The PIC is formed of the protein components of RNA PolII and the general transcription factors TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH (Schilbach *et al.* 2017, Lee & Young 2000). Mediator acts as a factor to stabilize the PIC, so that it may proceed into transcription (Kelleher *et al.* 1990, Allen & Taatjes 2015).

The polymerase C-Terminal Domain and its dynamic modification

The largest component of RNA Polymerase II, Rpb1, has a large disordered region at its C-Terminus made up of a number of repeats of the consensus amino acid sequence ($Y_1S_2P_3T_4S_5P_6S_7$). Metazoan Rpb1 constists of 26 repeats of this sequence in *S. cerevisiae* and 52 repeats in humans (Prelich 2002). The tyrosine, serine and threonine residues found in positions 1, 2, 4, 5 and 7 undergo reversible post-translational modifications through phosphorylation and de-phosphorylation. Polymerase exists with different com-

binations of phosphorylation sites to mark the different phases of transcription (Reviewed in Zaborowska *et al.* 2016 and Harlen & Churchman 2017).

Polymerase is phosphylated on the serine residue in position 5 (S_5) immediately following PIC assembly and the CTD is also shown to by phosphorylated on tyrosine 1 (Y_1) during early transcription in humans (Schlackow *et al.* 2017), however Y_1 is found to be phosphorylated towards the end of the gene in *S. cerevisiae* (Mayer *et al.* 2010). After initiation and moving into promoter escape, the serine residue in position 7 is phosphorylated, before these levels appear to decrease into the gene body. The levels of S_2 and threonine 4 (T_4) phosphorylation increase towards the end of the gene body and remain high after the poly(A) site into the termination region (Shown by ChIP-Nexus Harlen *et al.* 2016, also see Figure 1.3).

The TFIIH subunit Kin28 (Cdk7 in humans) phosphorylates both S_5 and S_7 (Akhtar *et al.* 2009) and is recruited as part of the capping complex to the 5' end of the RNA (Rodriguez *et al.* 2000). Depletion of Kin28 from *S. cerevisiae* cells results in accumulation of RNA PolII at the 5' ends of genes, suggesting phosphorylation of the CTD at the starts of genes assists in polymerase release from the promoter region (Rodríguez-Molina *et al.* 2016). In *S. cerevisiae*, the Cdk-homologs Bur1 (Cdk9 in humans) and Ctk1 (Cdk12 in humans) are the proposed kinases to phosphorylate S_2 (Bartkowiak *et al.* 2010, Pei *et al.* 2003). The phospho-kinases for the remaining CTD residues Y_1 and T_4 have not been determined in *S. cerevisiae*, however the role of the post-translational modifications can be investigated by other methods than disrupting or deleting the kinase. A recent study in a human cell line showed that by mutating the tyrosine residues in position 1 to a phenylalanine, which cannot be phosphorylated, this caused large amounts of transcriptional read-through downstream of the poly(A) site and termination defects (Shah *et al.* 2018). This is perhaps surprising given the location of phosphorylated Y_1 residues at the start of genes in human cells (Schlackow *et al.* 2017) impacting the polymerase at the end of the gene.

Post-translational modification of the polymerase CTD potentially serves as a direct method by which to demarcate the different phases of transcription. Phosphorylation of the CTD at different points could indicate the transitions of the polymerase from PIC

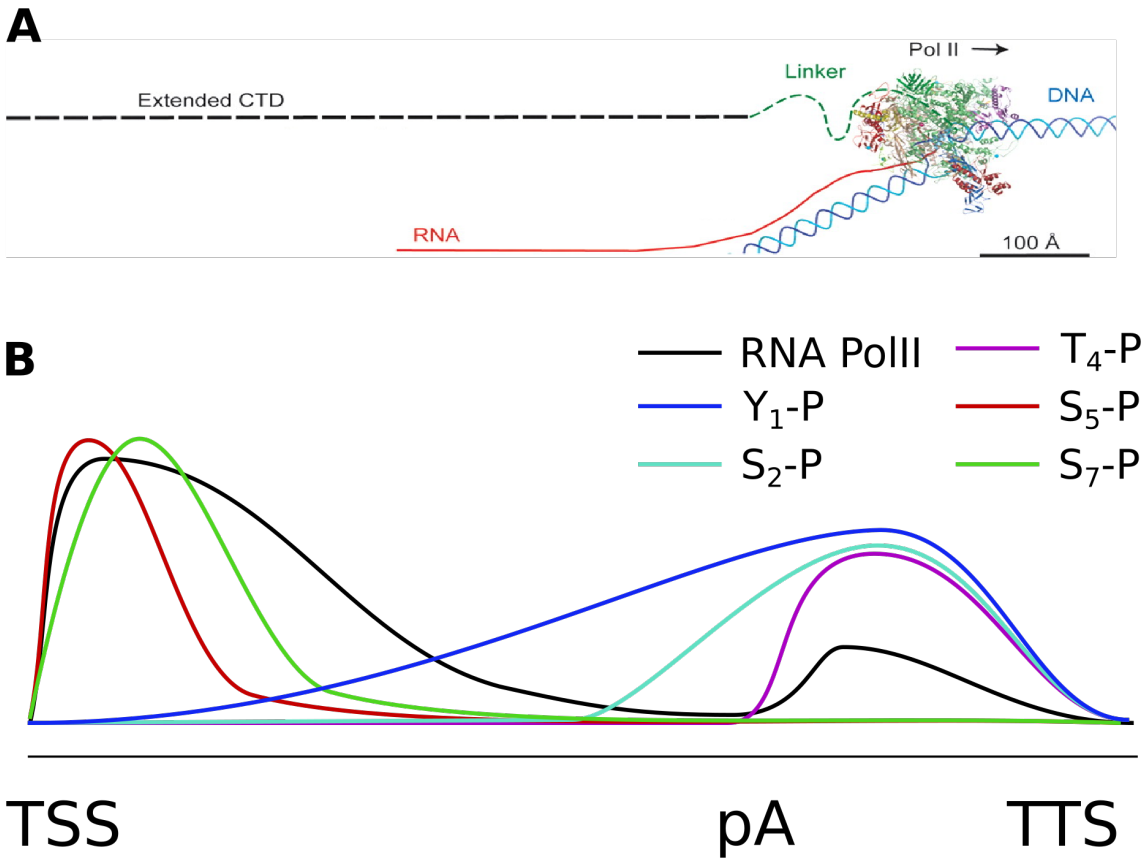


Figure 1.3: An outline of the post-translational modifications of RNA Polymerase II. (A) The polymerase molecule engaged with DNA with RNA molecule. The disordered C-Terminal Domain of Rpb1 is shown schematically, modified from Meinhart *et al.* 2005. (B) An outline of the average locations of polymerase and the post-translational modifications of the CTD. Tyrosine 1 (Y_1), serine 2 (S_2) and S_7 from ChIP-chip experiments in Mayer *et al.* 2010. Threonine 4 (T_4) and S_5 from ChIP-Nexus experiments in Harlen *et al.* 2016. PolII location from NET-seq experiment in Fischl *et al.* 2017. Levels are approximate with arbitrary scales and not normalised to total polymerase levels.

assembly to initiation, through promoter-proximal escape into the process of processive elongation. What is not known is to what extent each of these factors contribute to the dynamics of the polymerase through each of these phases or whether the process works in the opposite direction. Could disruption of the phosphorylation of the CTD result in the polymerase getting caught in the promoter, or being prematurely released from early elongation into the gene body?

1.3.2 Promoter escape

Once the PIC has formed and the serine in position 5 is phosphorylated, the polymerase complex proceeds from pre-initiation into an initiating state. The negative elongation

factors DSIF (DRB Sensitivity Inducing Factor) and NELF (Negative Elongation Factor) hold the polymerase in a position close to the promoter (Yamaguchi *et al.* 1999, Narita *et al.* 2003). The phospho-kinase Cdk9 (Bur1 in *S. cerevisiae*) phosphorylates the DSIF subunit Spt5 (hSpt5 in humans), converting DSIF from a negative to a positive elongation factor (Wada *et al.* 1998a, Wada *et al.* 1998b, Yamada *et al.* 2006) and releases NELF from the elongation complex (Sansó & Fisher 2013, Qiu & Gilmour 2017). Once NELF has dissociated from the complex, the Paf1 complex takes the place of NELF on the elongation complex, sitting on top of the core polymerase subunit and this allows TFIIS to join the RNA PolII complex (Vos *et al.* 2018a, Vos *et al.* 2018b, Adelman & Henriques 2018). The paused state of polymerase that is well reported in higher eukaryotes acts to inhibit further rounds of transcription (Shao & Zeitlinger 2017) as well as potentially act as a quality-control checkpoint (Laitem *et al.* 2015). Paf1 has been associated with both maintaining and releasing polymerase from its promoter-proximal paused state downstream of the TSS (Chen *et al.* 2015, Yu *et al.* 2015, Crickard *et al.* 2016). While the pause-site is often stated as occuring at a single site, it has been shown in *Drosophila* that polymerases may move at a slow speed following this initial pause site, then "dribbling" to a later location near the +1 nuclesome (Boija *et al.* 2017).

1.3.3 Transcription elongation

The main process of transcription is the conversion of DNA to complementary RNA, which occurs in the elongation step of transcription. Indeed, recent estimates suggest that elongation of typical human genes occurs on the timescale of tens of minutes, whereas initiation and promoter-proximal pausing has been shown to finish within a minute (Steurer *et al.* 2018). On this time-scale, we would expect transcription of a typical *S. cerevisiae* gene to last between 1 and 2 minutes, indicating much larger relative effect possible by affecting elongation or pause dynamics.

During the elongation stage, the polymerase must move through the chromatin template, assisted by the histone chaperone FACT and Spt6, which remove the histones from the DNA in front of the polymerase and replace histones after the polymerase has passed (Jeronimo & Robert 2016). Furthermore, the polymerase complex must overcome a num-

ber of blocks during transcription which cause stalling or pausing, such as proteins bound to the DNA or sequence-specific pause-sites (Larson *et al.* 2014). One method of polymerase release from a blocked state is by backtracking a small number of nucleotides in a 3'-5' direction, releasing the recently incorporated nucleotides from the active site of the polymerase. Backtrack release factor TFIIS (Dst1 in *S. cerevisiae*) assists in the cleavage of the 3' RNA extruding from the active site of the polymerase complex before it can proceed with processive elongation (Mason & Struhl 2005).

RNA splicing

In higher eukaryotes most genes are made up of coding regions (exons) and non-coding regions (3' UTRs and introns), which are removed from the RNA before translation may occur. In *S. cerevisiae* roughly 300 genes contain introns and it is suggested that the splicing out of these introns occurs co-transcriptionally, so that once termination has occurred, the mRNA is complete and ready for export and translation (Wallace & Beggs 2017). Transcription and splicing are coupled processes, with the speed of the polymerase governing the efficiency of splicing (Aslanzadeh *et al.* 2018) and the process of splicing affecting the speed of the polymerase, resulting in pausing (Nojima *et al.* 2015).

1.3.4 Transcription termination

Termination of protein-coding genes in eukaryotes is brought about by the cleavage and polyadenylation factor (CPF) binding to the nascent RNA and polymerase after transcribing the poly(A) sequence (Mischo & Proudfoot 2013). The read-through of the polymerase after the cleavage or poly(A) site has been shown by nascent transcription techniques to continue for up to 2kb in budding yeast (Baejen *et al.* 2017) and seen to continue well beyond 10kb in human cells (Laitem *et al.* 2015).

The torpedo model of termination

Originally proposed as a mechanism for degrading the cleaved RNA from the trailing polymerase (Connelly & Manley 1988), the torpedo model describes a method by which a 5' - 3' nuclear exonuclease degrades the RNA transcribed by the polymerase after cleavage

has occurred, then once the exonuclease reaches the polymerase, signals termination of transcription. This protein was detected as Xrn2 (Rat1 in *S. cerevisiae*), which when in a complex with Rai1 and Rtt103 binds the 5' end of the RNA and prompts the polymerase to begin termination (Kim *et al.* 2004, Baejen *et al.* 2017).

1.4 RNA processing

1.4.1 RNA export

Following processive transcription and efficient termination, RNA must be exported from the nucleus to the cytoplasm before it may undergo translation into functional protein (Figure 1.4). A number of proteins facilitate this process by binding to the elongation and termination complexes, forming protein-RNA complexes, travelling through the nuclear pore and deliver the RNA to the ribosomes (Reviewed in Katahira 2012). The coupling of transcription termination to RNA export is facilitated by the THO complex and TRanscription-EXport (TREX) complex (Aguilera 2005), which travels with the elongation complex before binding to the transcribed mRNA, forming an RNA-protein complex (ribonucleoprotein, or RNP). Mex67, a protein found on chromatin, bound to RNA and and the nuclear pores (Segref *et al.* 1997) binds to the RNA of actively transcribed genes and transports the RNA through the nuclear pores to the cytoplasm for translation (Dieppois *et al.* 2006).

Upon activation, genes have been shown to localise at the nuclear periphery, most likely to facilitate the rapid export of transcripts to the cytoplasm immediately after transcription has completed (Brickner *et al.* 2016). This may allow the tethering of actively transcribed genes to the nuclear pore complex, especially if following multiple rounds of transcription, termination and export factors remain at the site of transcription, maintaining contact with the nuclear periphery.

1.4.2 RNA degradation

All RNA in the cell must be degraded, recycling the total RNA and NTPs (Nucleo-TriPhosphates) present in the cell and, importantly, acting as a quality-control mecha-

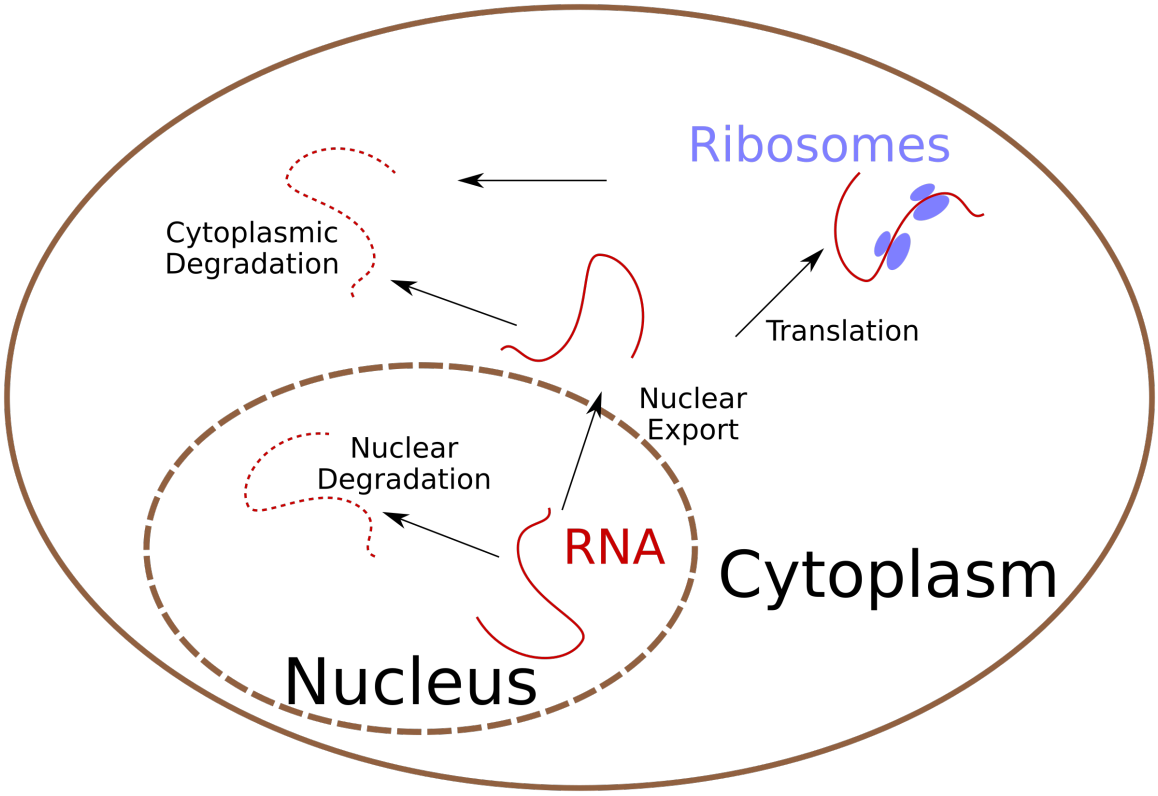


Figure 1.4: The pathways of RNA processing in the cell. Coding mRNA is exported from the nucleus after transcription, where it will undergo translation into the functional protein before being degraded. Non-coding RNA are targeted for early degradation either in the nucleus or the cytoplasm.

nism. Degradation of RNA will here be split into two categories, nuclear and cytoplasmic, although there are many complex pathways of RNA degradation (reviewed in Houseley & Tollervey 2009).

Nuclear degradation

Given the prevalence of non-coding transcription in eukaryotic organsims (Mellor *et al.* 2016), pathways exist to prevent the export of these transcripts and ensure quick turnover of these NTPs. These regulatory pathways also serve as a quality-control for coding transcription, assessing the efficacy of the transcription of genes to create the correct transcript for translation. The main enzymes involved in nuclear degradation are Rat1, as mentioned above, which degrades RNA from the 5' end (Connelly & Manley 1988, West *et al.* 2008) and Rrp6, a 3' to 5' exonuclease which targets unstable non-coding transcripts for degradation as well as forming further complexes to target a host of other RNA molecules (Makino *et al.* 2015).

Cytoplasmic degradation

Degradation of most RNA Pol II transcribed genes occurs either by the 5' to 3' exonuclease Xrn1, or via the exosome complex. Xrn1 has been earmarked as a key regulatory protein involved in a number of pathways closely associated with correct balance of RNA in the cell. As a quality-control enzyme, Xrn1 targets RNA which contain an early STOP codon, which if translated would form a truncated protein, via the Nonsense-Mediated Decay pathway (Wery *et al.* 2016). Furthermore, deletion of Xrn1 shows upregulation of a number of non-coding transcripts, termed XUTs as described above, highlighting the role of Xrn1 in removing spurious transcripts that reach the cytoplasm (van Dijk *et al.* 2011). More recently, Xrn1 has been slightly controversially termed a "synthegradon", affecting both the activation of gene expression as well as the degradation of RNA (Haimovich *et al.* 2013, Sun *et al.* 2013, Medina *et al.* 2014). Whether these effects are direct, with Xrn1 shuttling between the nucleus and the cytoplasm, or indirect via controlling the levels of transcription factors has not been determined, however it is clear that disruption of this cytoplasmic exonuclease affects not just the steady-state levels of transcripts in the cell, but also the ability of a cell to respond to external cues via the rate of synthesis of transcripts.

It is not known how exactly Xrn1 may facilitate gene activation and repression as well as monitor the mature RNA levels, or at what stage of transcription Xrn1 has the most profound effect. It is one aim of this study to determine where the effects of Xrn1 are most important in the process of RNA production, whether it be at the point of initiation, elongation or in assisting the passage of polymerase through the chromatin environment.

1.5 The dynamic nature of transcription

Many studies have focused on transcription and transcript production in populations of cells at steady-state. This has the potential to blur the true picture of how transcription is regulated as we perhaps do not expect every cell to always behave in the same manner. Indeed, across a population of cells, we expect cells to be in different phases of the cell cycle (Mischo *et al.* 2018), be undergoing different stages of metabolism (discussed below)

and vary through stochastic processes inherent to the cells (Thattai & van Oudenaarden 2001).

It is therefore necessary to take these factors into account when analysing population-averaged data, but where possible analysing cells either synchronised to any of the phases described above or analysing single cells presents a more powerful tool to look at the dynamics of gene expression. Cells do not exist in steady states, they are constantly adapting to external cues, turning over mRNA and proteins in the cell to make way for newly synthesised molecules and activate different regulatory networks. Studying cells moving through different biological states creates the opportunity to investigate the dynamics of the repression and activation of regulatory networks.

1.5.1 The Yeast Metabolic Cycle

S. cerevisiae cells have been shown to undergo metabolic oscillations under nutrient-limited conditions, in which a population of cells synchronise their consumption of oxygen (Tu *et al.* 2005) and metabolites in the media oscillate in a synchronous manner (Kuang *et al.* 2014, Reinke & Gatfield 2006). Furthermore, transcript levels of around 60% of genes in budding yeast also display oscillatory behaviour (Tu *et al.* 2005), with genes exhibiting peaks in expression throughout the cycle (Kuang *et al.* 2014, Nocetti & Whitehouse 2016).

The cycles exhibited by cells in the metabolic cycle range from as short as 40 minutes to many hours depending on the feeding rate of the cells. This creates a system under which dynamic gene expression can be studied, with many genes exhibiting rapid activation and repression. The coupling of regulatory factors such as the placement of nucleosomes, post-translational modifications of histones and transcription factors to gene expression has traditionally been performed via steady-state measurements, which reveal correlative interactions, without any temporal resolution.

The placement of the nucleosomes around the NDR at the promoter has been shown to correlate with transcription at a small number of genes during the metabolic cycle (Nocetti & Whitehouse 2016). However, given only ∼50 genes have their activation shown as occurring contemporaneously with a shift in the nucleosome immediately downstream of the TSS (+1 nucleosome), it suggests that there are further mechanisms governing the

rapid activation and repression of genes through the cycle.

In this study we will use this system to investigate how gene activity affects the activity of the polymerase whilst transcribing through the gene after initiation. Here we are presented with a system which exhibits natural fluctuations in gene expression, currently missing is how activation or repression of a gene affects the speed of a polymerase through chromatin. By studying the distributions of polymerase as overall levels fluctuate, we will be able to establish whether phenomena such as pausing, backtracking or early termination are affected by expression levels, or are inherent to the genes regardless of number or frequency of initiation events.

1.5.2 Gene expression under different carbon sources

S. cerevisiae cells require different enzymes to metabolise the carbon sources found in different sugars such as glucose and galactose. As such, alternative carbon sources create different gene expression profiles in *S. cerevisiae*. If yeast cells are moved from normal lab-based carbon source glucose to galactose this results in the induction of the *GAL* genes required for galactose regulation (Venturelli *et al.* 2012, Escalante-Chong *et al.* 2015). Galactose is taken up into the cells by Gal2 (Tschopp *et al.* 1986), which activates phosphorylation of Gal3 which then represses the *GAL* system repressor Gal80 (Torchia *et al.* 1984). The transcription factor Gal4 is usually repressed by Gal80, however after this repression is removed, Gal4 activates the transcription of a number of genes, namely *GAL2* and *GAL3*, creating a positive feedback loop, *GAL80* and the *GAL* locus (Venturelli *et al.* 2012 and see Figure 1.5).

1.5.3 The *GAL* locus

At the centre of the switch in metabolism from glucose to galactose is the heavily transcribed *GAL* locus, consisting of galactose metabolism genes *GAL1*, *GAL7* and *GAL10* (See Figure 1.6). These genes are regulated by the transcription factors Gal4 and Mig1 (Nehlin *et al.* 1991), which bind upstream of the transcription units. In glucose-containing media however, both *GAL1* and *GAL10* exhibit high levels of antisense transcription reading the reverse-strand of each gene (Figure 1.6) with the well-studied long-noncoding

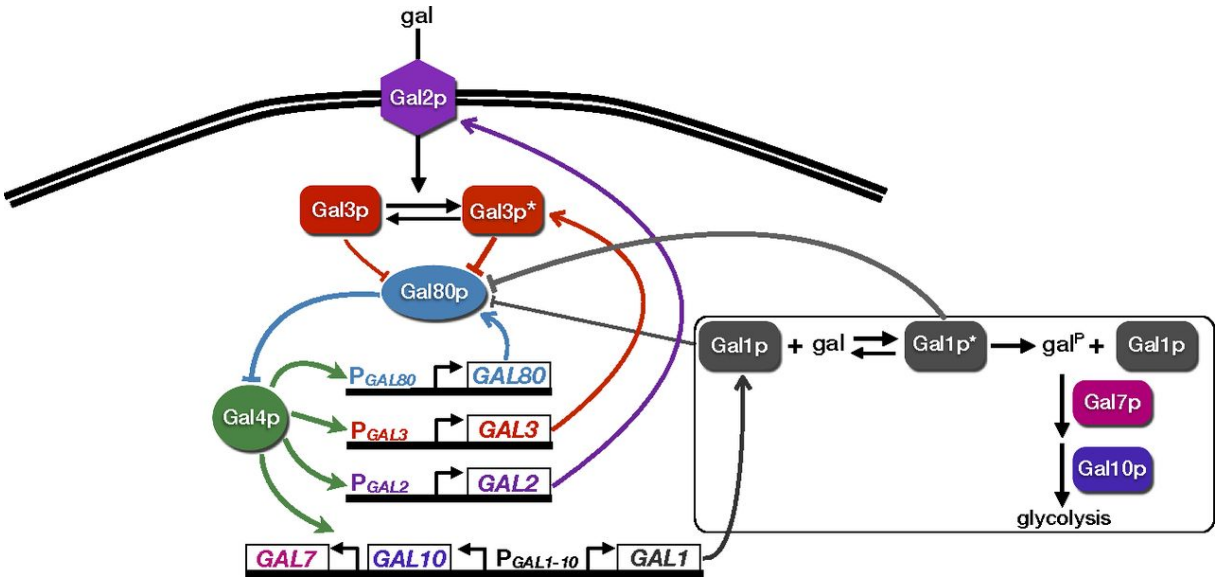


Figure 1.5: A schematic showing the cascade of protein interactions which initiate galactose metabolism. Shown are the proteins which are activated upon detection of galactose and the activation of the *GAL* locus. Modified from Venturelli *et al.* 2012.

GAL10 antisense regulated by transcription factor Reb1 (Houseley *et al.* 2008, Lenstra *et al.* 2015). It was found by placing a strand-specific block to the transcription of the *GAL10* lncRNA in Lenstra *et al.* 2015 that the lack of this antisense transcription through the shared *GAL1-10* promoter that in typically repressed conditions in glucose both *GAL1* and *GAL10* exhibited leakage of transcription, potentially offering one important role for this antisense transcription unit. Upon a change in media from glucose or raffinose to galactose, the *GAL10*-ncRNA is switched off, allowing expression of both the sense *GAL10* RNA and the *GAL1* RNA, both expressed via their shared promoter. Activation of *GAL1* also sees the shutting off of the *GAL1* antisense transcription, indicative of the observed exclusive expression of either sense or antisense at a single locus (Nguyen *et al.* 2014). It has been further suggested that the expression of this long *GAL10*-ncRNA may perform a regulatory role by forming RNA-DNA hybrids, known as R-loops, where the transcribed RNA pairs with the exposed complementary DNA strand in the transcriptional bubble (Cloutier *et al.* 2013, Cloutier *et al.* 2016).

It is currently not known how the transcription of the *GAL1* antisense transcript affects the transcription of the corresponding sense transcript. Given there is a lot of overlap between the two transcription units, we may expect a lot of interplay between the two, perhaps by the antisense transcription unit overlapping the sense promoter and repressing sense expression (Eser *et al.* 2016) or by creating a more activating chromatin

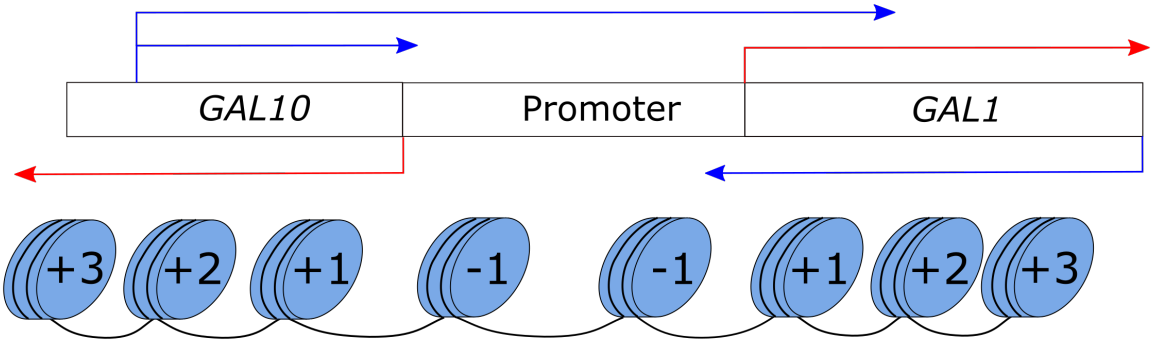


Figure 1.6: The layout of transcription units at the *GAL* locus. Shown are the location and orientation of the galactose metabolism genes, sense transcription units are in red and antisense transcription units are in blue. Nucleosome positions shown are taken from Bryant *et al.* 2008.

template (Murray *et al.* 2015). Certainly very little is known about the rate at which the polymerase undergoes initiation and elongation through this crowded locus and it may not be as simple as a case of antisense causing repression or activation of the sense transcription.

1.6 Techniques used to study transcription dynamics

Advances have been rapid in the technology available to study transcription directly, as opposed to indirectly via RNA or protein levels. The development of these techniques generally either rely on studying the locations of RNA PolII when bound to DNA, or the nascent RNA while still part of the elongation complex. Given each technique focuses on different aspects of transcription, it is perhaps unsurprising that each reveal different aspects about transcription as a process, however this has created a complex picture of transcription.

Outlined here are a number of techniques with single-gene and genome-wide focuses which each contribute to the study of transcription and, in particular, transcription dynamics.

1.6.1 Single-gene analysis

RNA FISH

Traditional methods of RNA quantification and localization have relied on population-wide techniques, assessing global changes in levels of a given molecule by measuring the levels present in all cells in a population and inferring the same must be true at the level of the individual cell. Northern blotting probes for an RNA sequence of interest, showing absolute levels in a population of cells, however this hides any further information regarding the distribution of RNA across these cells. Similar problems arise through high-throughput sequencing of the RNA molecules in a population of cells, where the total levels of RNA seen at each gene may equally be reflective of equal expression levels in all cells, or a large range in expression levels across a population.

RNA fluorescence *in-situ* hybridization (FISH) (Raj *et al.* 2006, Zenklusen *et al.* 2008) adds fluorescently-labelled single-stranded DNA probes, complementary to the RNA sequence of interest, into cells to assess the number and position of RNA molecules of a particular gene in a cell and across a population. The most immediate observation using this technique is that the distribution of RNA produced by a single gene is not homogeneous across a population of identical cells (Raj *et al.* 2006) and cells with high numbers of RNA molecules balancing with cells barely expressing a gene contributing to the population-wide averages seen by blotting and other techniques. Furthermore, RNA FISH experiments performed at a number of loci in *S. cerevisiae* highlighted that the distributions of RNA across a population of cells, both nuclear and cytoplasmic, are distinct for different genes (Zenklusen *et al.* 2008). Adopting modelling techniques designed for analysing the production of proteins (Peccoud & Ycart 1995) to the synthesis and degradation of RNA highlighted the non-Poissonian dynamics of RNA production and has brought into question the idea of all genes undergoing constitutive expression, either existing in an active or inactive state. It is likely that many genes exist in "bursty" states when active, with a promoter existing in two or more states when active, resulting in a range of levels of promoter activity (Zenklusen *et al.* 2008, Corrigan *et al.* 2016). In particular Zenklusen *et al.* 2008 use the bursty model proposed by Peccoud & Ycart 1995 to analyse 5 genes, concluding that one of the five genes studied displayed bursting

behaviour distinct from the constitutive 'telegraph' model. Modelling the nuclear and cytoplasmic distributions of RNA produced from a single gene has been able to determine the rates of transcription initiation events, the rate of polymerase elongation (Zenklusen *et al.* 2008) and the stability of RNA (Trcek *et al.* 2011).

RNA PolII ChIP on a single gene

One particularly powerful tool developed to study the elongation rates of actively transcribing polymerase in yeast, is to look at the levels of polymerase bound to different sites of a long gene under activation. A number of studies (e.g. Malik *et al.* 2017) have utilised the galactose-inducible GAL-promoter to initiate transcription across a template which is turned off in glucose-containing media and rapidly activated upon changing to galactose-containing media. By placing this promoter upstream of a long gene (in the case of Malik *et al.* 2017, the ∼8kb gene *FMP27*) and probing for RNA PolII at different sections of the gene, it is possible to track the speed of polymerase across this template both upon activation (entering wave) and after repression (clearing wave). These experiments have resulted in estimates of RNA PolII elongation speed between 1.5kb/min and 2kb/min for these genes. One caveat with these techniques is that these measurements are taken under changing conditions for the cell and one must assume that the catalytic activity of the polymerase molecule is not affected by these changes and reflect the kinetics at steady-state at this gene. Another restrictive assumption is that the dynamics of polymerase remains the same at every gene. In order to obtain a full map of polymerase dynamic ability, these experiments would need to be performed through a range of chromatin environments, under different external conditions and through different lengths of genes to truly obtain a representative picture of the dynamic range of elongating RNA PolII.

There are also a number of technical restrictions associated with studying polymerase dynamics in this manner. The resolution of the technique is also restricted by the size of DNA fragments which are associated with the chromatin and RNA PolII, typically in the region of 150-200bp. This means that phenomena such as stalling or backtracking on any scale less than this will be conflated with the surrounding region of the DNA.

Furthermore, this technique requires cross-linking of the polymerase complex to the DNA or chromatin, often via UV-irradiating the cells or adding formaldehyde. This can create stress-responses in the cell, for example DNA damage brought about by exposing cells to UV and the corresponding response (Birrell *et al.* 2002, Fu *et al.* 2008). Care must therefore be taken interpreting these results given the stress the cells have been placed under.

1.6.2 Genome-wide analysis

High-throughput sequencing techniques

Chromatin immuno-precipitation followed by high-throughput sequencing (ChIP-seq Johnson *et al.* 2007, ChIP-exo Rhee & Pugh 2011 or ChIP-Nexus Harlen *et al.* 2016) allows the localization and quantification of a particular protein bound to the chromatin-bound DNA. To study nascent transcription, subunits of RNA PolII are pulled-down, showing the regions of the genome bound by polymerase and, presumably, transcribing. This is a useful technique to obtain a proxy for transcriptional levels at each gene, however there are two main drawbacks of this technique. Firstly, by extracting proteins bound to chromatin, we are unable to distinguish which strand of the genome the protein is targeting, a particular problem when studying transcription as the polymerase is known to transcribe both the sense and antisense strands of a gene and this signal will be a conflation of the two. Secondly, the resolution of this technique is dependent on the length of DNA cross-linking to the chromatin and protein of interest, typically in the region of 200bp. This becomes a problem when attempting to distinguish the location of polymerase to the promoter, pausing site or gene body, which have been suggested to exist in different dynamic regimes (Steurer *et al.* 2018).

A number of techniques have been developed to combat the problems highlighted with ChIP-seq, which have mainly relied upon a run-on of elongation-competent polymerase, incorporating labelled nucleotides into the RNA sequence. First developed was Global Run-On sequencing (GRO-seq, Core *et al.* 2012), in which nuclei from cells are isolated, then labelled NTPs are added to the nuclei, where the polymerases undergo an *in vitro* transcription run-on. This technique has the advantage of giving strand-specific locations

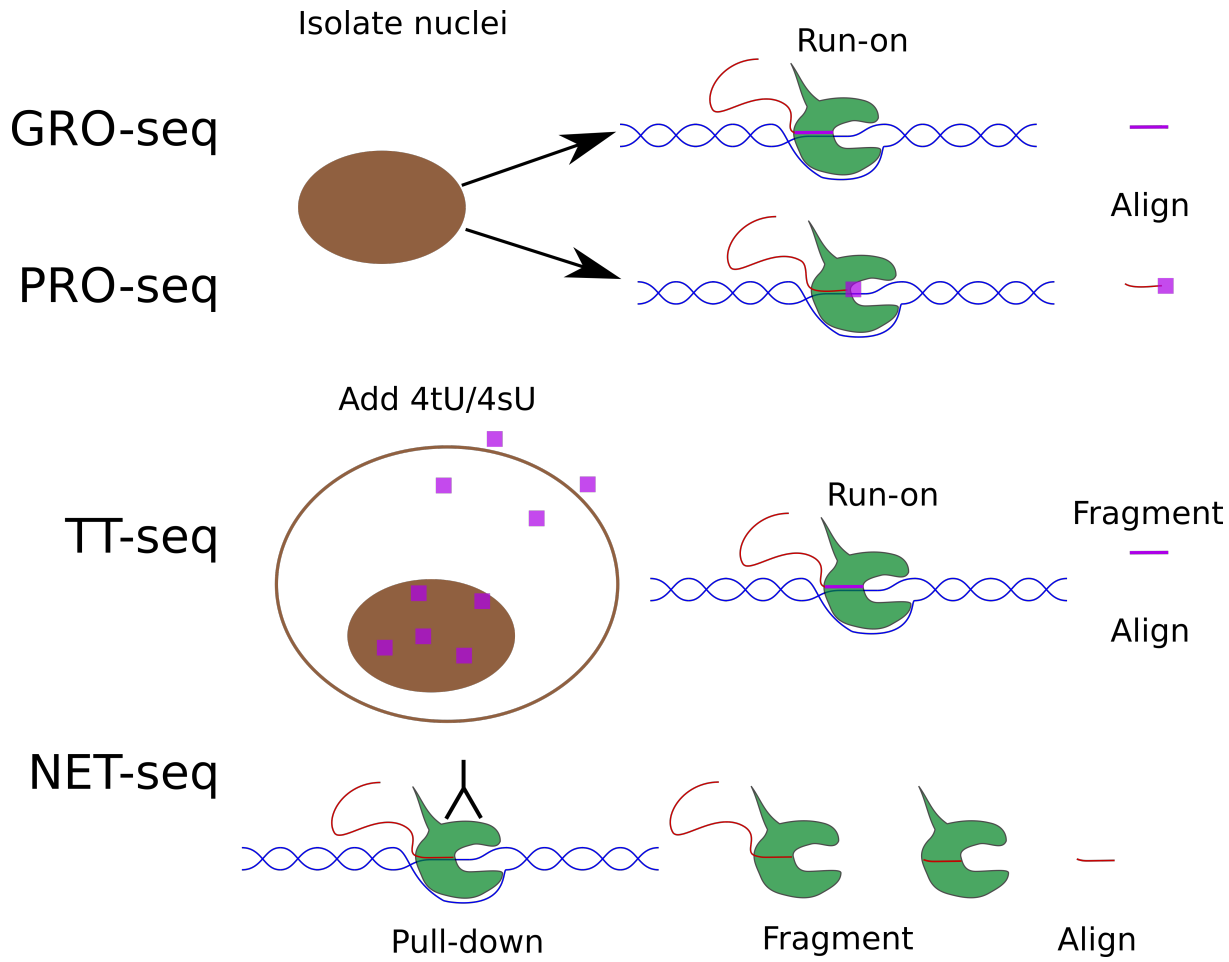


Figure 1.7: Outline of approaches to study polymerase activity. In GRO-seq and PRO-seq, nuclei are isolated before an *in vitro* run-on reaction is performed with labelled nucleotides. In PRO-seq, a single nucleotide is incorporated to the 3' end of the RNA, which is taken forward for sequencing. Under TT-seq, alternative NTPs, 4tU or 4sU are added to the cells and then incorporated into the nascent RNA. RNA is then fragmented before sequencing. In NET-seq, the flag-tagged RNA PolII molecule is pulled down, the nascent RNA in the active site of the polymerase is extracted and taken forward for sequencing.

of the polymerases which undergo this *in vitro* elongation. In order to block elongation and release polymerase from promoter-proximal pausing, sarkosyl is added creating un-physiological conditions in the cell, in particular allowing for the release of paused polymerases into elongation (Rougvie & Lis 1988). This technique, however, does not resolve the issue of low resolution of the steady-state positions of polymerase, as the length of labelled sequence produced by the polymerase is proportional to the length of run-on reaction, typically around 5 minutes, resulting in a resolution of many kb.

To increase the resolution of the run-on sequencing, Precision Run-On sequencing (PRO-seq) was developed, under which a single labelled nucleotide is incorporated during the run-on reaction (Kwak *et al.* 2013, Booth *et al.* 2016). This is a very powerful

method for determining the location of polymerase in relation to the separate regions of transcription units and in particular is not sensitive to those non-coding transcripts which undergo rapid degradation. However this still relies on an *in vitro* elongation reaction, therefore elongation-incompetent polymerases will still undergo the elongation step.

As a more native alternative approach to the *in vitro* run-on techniques described above, Transient-Transcriptome sequencing (TT-seq) has been developed in human cells and in *S. pombe* (Schwalb *et al.* 2016, Eser *et al.* 2016). In TT-seq, 4-thiouracil (4tU) or 4-thiouridine (4sU) is taken up by the cells and incorporated into the RNA as an alternative to UTP. After fragmentation of the sequences, those with these modified UTPs are taken forward for sequencing and give an indication of the levels of nascent transcription occurring at each transcription unit during a labeling window, typically 5 minutes again giving low-resolution reads. However, the introduction of these chemicals to cells results in aberrant phenotypes (Burger *et al.* 2013) or is toxic to the cells. As a result, interpretation of these results must be performed with care.

Finally, as a more direct comparison for RNA PolII ChIP-seq, NET-seq and mNET-seq have been developed to study the locations of transcriptionally-engaged polymerases in a strand-specific high-resolution manner (Churchman & Weissman 2011, Nojima *et al.* 2015). In NET-seq, a sub-unit of RNA PolII is FLAG-tagged, then the FLAG-tag is pulled down, bringing the polymerase complex and the associated RNA. The RNA found in the active-site of the polymerase complex is then isolated and sequenced, the most 3' nucleotide is the last nucleotide to be incorporated into the polymerase and thus gives the location of the polymerase on the DNA.

The development of these high-resolution quantitative techniques offer a fantastic opportunity to construct models which in turn will help us understand the intricacies of the underlying mechanisms of transcription and gene regulation. Techniques such as RNA PolII ChIP-seq and NET-seq have highlighted the non-uniform distributions of transcriptionally-engaged polymerase complexes across genes, though it is still unknown to what extent these distributions point to function and how these arise from the underlying dynamics (Churchman & Weissman 2011, Ehrensberger *et al.* 2013). As part of this work, we aim to use NET-seq data to construct mathematical models of transcription,

indicating the key interactions and steps in transcription, and then use these models to probe the roles of individual factors in establishing polymerase dynamics. Comparing simulation results with wild-type NET-seq data and following genetic intervention, we will begin to uncover the effects of individual components of the transcription network in establishing these dynamic regimes.

RNA PolII FRAP

The recovery of fluorescence after photo-bleaching has been used on a genome-wide scale to study the diffusion rates of a number of proteins (Mazza *et al.* 2007, Mueller *et al.* 2010, Steurer *et al.* 2018). Here, we focus in particular on the photo-recovery of GFP-tagged RNA PolII. By tagging the polymerase molecules in the cell, then photo-bleaching, it is possible to detect the rate of diffusion of the molecules *in-vivo* and at single-cell resolution.

Analysis of the recovery curves of the fluors after photo-bleaching has highlighted the need for polymerase to exist in multiple dynamic states, with sub-populations of polymerases exhibiting altered recovery rates through diffusion (Steurer *et al.* 2018, Darzacq *et al.* 2007). These various states of polymerase have been interpreted as freely-diffusing, initiating, promoter-proximal paused and elongating polymerase, each with their own inferred rates of diffusion, relating to the photo-recovery.

While this approach offers the possibility to determine the rates of elongation, polymerase diffusion and release from static states whilst transcribing in individual cells, it must do so by treating genes as homogeneous entities. By photo-bleaching half of the nucleus, a whole spectrum of transcriptional activity is encompassed with this technique, including genes that may or may not undergo promoter-proximal pausing, long genes producing functional transcripts, short genes including those producing non-coding PROMPTS or eRNAs and all possible chromatin landscapes. The true power of this tool has perhaps not yet been realised, but currently it serves to muddy the true underlying dynamics of transcription and should perhaps be restricted to analysis of individual loci for the time being.

There are a number of studies that use FRAP on a single gene (Corrigan *et al.* 2016, Maiuri *et al.* 2011). One immediate conclusion from these papers in contrast to genome-

wide FRAP studies, is that depending on the gene studied, one can obtain transcription rates that differ wildly from the genome-wide averages obtained through other techniques. While this is a powerful technique for understanding the dynamics of RNA PolII, one must be careful interpreting results on the single-gene and genome-wide scale in the context of each other.

1.7 Factors affecting transcription and gene regulation

The question of how transcription and elongation factors affect transcription and gene expression has become more nuanced with the advent of new experimental techniques. A traditional view of gene expression would be that transcription factors turn genes on and off upon environmental cues, as judged by analysing RNA or protein levels at individual genes (Northern or Western blotting) or via genome-wide analysis (RNA-seq or proteomics). For a number of transcription factors, this does appear to be the case, for example Rap1 is a transcription factor known to be present at the promoters of Ribosomal-Protein Genes (RPGs), which leaves the promoters upon environmental stress and is accompanied by a drop in transcript levels (Knight *et al.* 2014). However, a recent study found that Rap1 also assists in blocking cryptic non-coding transcription at a number of genes, and in particular blocks spurious transcription that reads across coding regions of the genome, but that do not encode for protein-coding RNA (Candelli *et al.* 2018). This complicates matters as we can no longer view transcriptional regulation as a merely binary process of on/off or even factors only having a single role in the cell.

The advent of nascent transcription techniques such as PRO-seq, GRO-seq or NET-seq show that transcription and elongation factors can affect not just the distribution of actively transcribing polymerases between genes, but the distribution of polymerase across a single gene (e.g. Fischl *et al.* 2017, Chen *et al.* 2015, Churchman & Weissman 2011). This has instigated discussion around the role of factors in the journey of a polymerase through a gene and how this contributes to the expression of individual genes after the supposed on/off switch at the promoter. As such the list of proposed factors affecting gene

expression and the myriad ways they impact transcription is ever-growing. A number of these factors are discussed here:

1.7.1 Transcription and elongation factors

The recruitment of the polymerase and the PIC is performed by a number of factors which target the promoter regions of genes. The TATA-box or "TATA-like" sequences found in the promoters are target-sites for the general transcription factors, in particular the TATA-binding protein TBP (Rhee & Pugh 2012) as part of TFIID or the SAGA complex. While these general transcription factors perform obvious activating roles, other transcription factors also act to repress the activity of genes, or even the same factor acts to both activate and repress genes under different circumstances (Scully *et al.* 2000, Latchman 2001, Shore & Nasmyth 1987).

Once the polymerase complex has proceeded beyond the stage of initiation, a number of further factors are recruited to polymerase either directly, or by binding the nascent RNA (Battaglia *et al.* 2017, Fischl *et al.* 2017). TFIIS (Dst1 in *S. cerevisiae*) has been shown to alleviate backtracking events and increase the processivity of the polymerase complex (Mason & Struhl 2005, Sigurdsson *et al.* 2010). Other elongation factors have specific roles such as assisting the polymerase through the chromatin template such as FACT (Belotserkovskaya *et al.* 2003) or binding the polymerase complex in the gene body, acting to influence the dynamic state of the polymerase by releasing or causing pausing as has been suggested for Paf1 (Chen *et al.* 2015, Yu *et al.* 2015, Fischl *et al.* 2017).

Many factors bind the elongation complex at different points of the transcription process, performing roles specific to initiation, promoter-proximal region release, elongation and termination. These factors may not influence gene expression in a binary fashion like the transcription factors of the RPGs as discussed above, but rather fine-tune the levels and timing of the production of RNA molecules from transcription events.

1.7.2 DNA and chromatin

Transcription occurs across the DNA through the context of the chromatin environment. The histones which make up the nucleosomes undergo a large degree of post-translational

modifications, with the histone tails exhibiting extensive methylation, acetylation, ubiquitination and other modifications. While the DNA sequence dictates the region of interest to be transcribed and directs the landing sites of transcription factors and the PIC, there are further downstream obstacles to the process of transcription.

The methylation of H3K4 and H3K36 are typically associated with active transcription of genes (Santos-Rosa *et al.* 2002, Howe *et al.* 2017), however recent work has suggested that the placement of these marks are likely to be a result of the act of transcription (Kuang *et al.* 2014, Soares *et al.* 2017). Methylated histone residues also recruit further factors, such as the methylation of H3K36 by Set2, which recruits the histone de-acetylase complex (HDAC) Rpd3S to the chromatin. The levels of acetylation on the chromatin have been shown to modulate the speed of the polymerase (Brown *et al.* 2018), suggesting both direct effects of acetylation levels and indirect effects on the levels of histone methylation.

1.7.3 Non-coding transcription

In higher eukaryotes, enhancer elements contribute to the expression of protein-coding genes. These elements can be found within the coding region of the corresponding gene or even many kb from the target gene (Zacharioudakis & Tzamarias 2017). *S. cerevisiae*, however does not have annotated enhancer regions but does have a rich landscape of non-coding transcription, including upstream activating sequences (UAS, Marchler *et al.* 1993, Dobi & Winston 2007) or upstream regulating sites (URS, Ozcan *et al.* 1997). There have been a number of roles suggested for non-coding transcription or the RNA, such as preventing transcriptional leakage at sense promoters (Lenstra *et al.* 2015), repressing transcription on the opposite strand (Eser *et al.* 2016) or changing the chromatin landscape via nucleosome placement and histone post-translational modifications (Murray *et al.* 2015, Brown *et al.* 2018). One issue with attempting to find a universal role of non-coding transcription and particularly antisense transcription is that many studies are performed on single genes or loci, with distinct chromatin architecture, regulatory factors and surrounding genes (Moretto *et al.* 2018, Ietswaart *et al.* 2017). In order to address this, we must combine observations made at individual loci with bioinformatic trends on

a genome-wide scale whilst keeping in mind the distinct nature of each gene regarding chromatin, 3D-architecture and regulatory factors. Ideally, one would wish to globally alter the levels of antisense transcription in the same manner to truly observe the effects on sense transcription, to determine whether effects are ubiquitous, or locus-specific.

1.7.4 RNA processing factors

The traditional metric of gene expression has been to measure the amount of RNA produced from each gene. As RNA levels are a product of both transcription and degradation, depleting nuclear or cytoplasmic degradation factors will change the levels of RNA and particularly non-coding RNA which are targeted for rapid degradation (van Dijk *et al.* 2011). However, use of 4tU-labelling in *xrn1*Δ strains showed that both the synthesis and the degradation of RNA are affected, suggesting multiple roles for this particular exonuclease (Sun *et al.* 2013). The same study, however, showed that the same was not true for an *rrp6*Δ strain, suggesting that any disruptions to the levels of RNA after the loss of the nuclear exonuclease is compensated for by another buffering mechanism.

A further important role is that of export factors which either facilitate the export of protein-coding RNA or retain the non-coding transcripts. Paf1 has been shown to not only have a role in elongation (Yu *et al.* 2015), but also to mark elongation complexes transcribing protein-coding genes to be exported (Fischl *et al.* 2017). If nuclear export is deficient, this gives another step in the gene expression stage where RNA is efficiently produced, but will not be translated into functional protein.

1.8 Distributions of quantitative data reveal more information than summary statistics

While much can be determined from simple statistics such as the mean and variance of a distribution, this at times hides the true information and can be misleading. As has been highlighted in Munsky *et al.* 2018, if the assumptions of normality are violated, we cannot apply the central limit theorem to make sensible conclusions when comparing two datasets. Indeed, a change in the mean levels of, for example, a particular RNA produced

between two conditions may be due to consistent increases in the production of RNA in every cell, or the balance changing between two sub-populations where one may have high production levels and the other low. Distributions such as those with long tails or which are bimodal violate the Central Limit Theorem, therefore we must be more careful with the interpretation of these results.

1.8.1 Constructing models to obtain distributions

Given the non-Normal distributions often found in biological scenarios, (long-tailed distributions of RNA across a population as in Zenklusen *et al.* 2008, bimodal distributions of yeast cell-size as in Tu *et al.* 2005), we wish to determine metrics which best describe these distributions and, in particular, describe changes in these distributions in a quantitative manner. The distributions presented here are derived from biological contexts, taking known or speculated interactions from the literature which may contribute to the output measured in each context. There may be many models which fit a given dataset and one goal here is to determine a set of minimal reactions required to replicate the distributions analysed or utilise existing models from the literature (e.g. Raj *et al.* 2006, Choubey *et al.* 2015).

To determine the "minimal" or best-fit model required to reproduce the distributions from the raw data, we use information criteria to assess the fit of models to data. As set out in Burnham & Anderson 2004, the Akaike Information Criteria or Bayesian Information Criteria determine the likelihood of observing the data from a given model, whilst penalising for the complexity of the model. These are given by:

$$AIC = 2k - 2ln\left(\hat{L}\right)$$
$$BIC = ln(n)k - 2ln\left(\hat{L}\right)$$

where k is the number of free parameters, or complexity of the model, n is the number of datapoints fit to the distributions and $\hat{L}$ is the maximum likelihood of the data under the given model. When comparing two models, the evidence supporting the value with smaller AIC can be interpreted through the absolute differences between the AIC values

as given in Burnham & Anderson 2004, here called Δ_{AIC}:

	Support for model with larger AIC
$\Delta_{AIC} \leq 2$	Substantial support
$4 \leq \Delta_{AIC} \leq 7$	Considerably less support
$\Delta_{AIC} > 10$	Essentially no support

Once a best-fit model has been determined, we can infer the likely underlying structure of the biological system of interest, and use this to predict the biological outcome from perturbing the system. After performing the given biological perturbation and observing the outcome, we can then use this to assess the predictive power, and thus the relevance, of the model and alter as appropriate.

The constructions of models goes beyond merely replicating the distributions of *in vivo* data. We can use these models to obtain numerical estimates on the changes in dynamic rates of a system under perturbation. It is the main aim of the work presented here to explain the dynamic ranges of the steps of gene expression and transcription and the extent to which these rates are altered under genetic intervention. This will in turn assess the role of key factors in gene expression and how they contribute to the dynamics of the system.

1.8.2 Existing models of transcription

Many transcription models have been developed to fit a specific purpose. The use of single-molecule RNA FISH to count and localise the RNA from a particular locus has been followed with use of stochastic models originally designed to analyse the distributions of proteins across a population of cells (Peccoud & Ycart 1995). By taking similar stochastic models of synthesis, studies have used bursting promoter kinetics as a key feature in models which output the number of nascent and cytoplasmic RNA per cell (Zenklusen *et al.* 2008, Raj *et al.* 2006). These modelling approaches have outlined the heterogeneity of transcription dynamics across genes in the same organisms. In particular, Zenklusen *et al.* 2008 found that of the 5 genes analysed, at least one gene appeared to undergo transcriptional bursting, as opposed to constitutive expression at others.

As more nascent RNA techniques were developed, further models were required to understand the complexity of the underlying kinetics that contributed to these data. Choubey *et al.* 2015 showed that the distributions of nascent RNA produced by techniques such as RNA FISH were strongly supportive of a multi-state polymerase, with altered catalytic activity in each state to proceed along the DNA template. Further complexity has been added with the use of live-cell imaging of fluorescently tagged RNA, offering live-cell imaging of nascent transcription dynamics. In *Dictyostelium* cells, Corrigan *et al.* 2016 reported that the dynamics observed were best fit by a promoter not necessarily undergoing switch-like bursting behaviour, but by a promoter which exists in a continuum of states, varying from always active to always inactive.

Looking beyond RNA as the output of transcription, RNA PolII ChIP experiments offer a view of the activity of polymerase and the distributions of transcriptionally-engaged RNA PolII. Ehrensberger *et al.* 2013 argued by using simple models of polymerase advancement that promoter-proximal pausing is indistinguishable from a high initiation rate when looking at large peaks at the 5' ends of genes via ChIP. This has opened the field to more theoretical models of RNA PolII transcription, whereby stalling, backtracking and degradation of polymerase take central roles in the action of converting DNA to RNA through a complex enviromnent (Li & Zhang 2015).

1.9 Data included and analysed in this study

Organism	Background	Experimental data	Source
S. cerevisiae	BY4741	NET-seq - glucose	Fischl *et al.* 2017
S. cerevisiae	BY4741	NET-seq - galactose 5 minutes	Fischl *et al.* 2017
S. cerevisiae	BY4741	NET-seq - galactose 15 minutes	Fischl *et al.* 2017
S. cerevisiae	BY4741	NET-seq - galactose 60 minutes	Fischl *et al.* 2017
S. cerevisiae	BY4741	RNA-seq - glucose	This study, performed by Harry Fischl
S. cerevisiae	BY4741	RNA-seq - galactose 5 minutes	This study, performed by Harry Fischl
S. cerevisiae	BY4741	RNA-seq - galactose 15 minutes	This study, performed by Harry Fischl
S. cerevisiae	BY4741	RNA-seq - galactose 60 minutes	This study, performed by Harry Fischl
S. cerevisiae	BY4741	MNase-seq - glucose	This study, performed by Ronja Woloszczuk
S. cerevisiae	BY4741	MNase-seq - galactose 15 minutes	This study, performed by Ronja Woloszczuk
S. cerevisiae	BY4741	MNase-seq - galactose 60 minutes	This study, performed by Ronja Woloszczuk
S. cerevisiae	BY4741 - *GAL1* - ADH terminator	RNA FISH, Northern blot	Brown *et al.* 2018
S. cerevisiae	BY4741 - *GAL1* - scrambled ADH terminator	RNA FISH, Northern blot	Brown *et al.* 2018
S. cerevisiae	*set3*Δ - *GAL1* - ADH terminator	RNA FISH, Northern blot	Brown *et al.* 2018
S. cerevisiae	*set3*Δ - *GAL1* - scrambled ADH terminator	RNA FISH, Northern blot	Brown *et al.* 2018
S. cerevisiae	BY4741 *ARG4* - GFP tagged	RNA FISH	This study, performed by Meredith Wouters
S. cerevisiae	BY4741 *ATP4* - GFP tagged	RNA FISH	Brown *et al.* 2018
S. cerevisiae	BY4741 *BAT2* - GFP tagged	RNA FISH	This study, performed by Meredith Wouters
S. cerevisiae	BY4741 *GCV3* - GFP tagged	RNA FISH	Brown *et al.* 2018
S. cerevisiae	BY4741 *GDH1* - GFP tagged	RNA FISH	This study, performed by Meredith Wouters

Organism	Background	Experimental data	Source
S. cerevisiae	BY4741 *HMS2*-GFP tagged	RNA FISH	Brown *et al.* 2018
S. cerevisiae	BY4741 *HOM3* - GFP tagged	RNA FISH	This study, performed by Meredith Wouters
S. cerevisiae	BY4741 *KAP104* - GFP tagged	RNA FISH	This study, performed by Meredith Wouters
S. cerevisiae	BY4741 *PDA1* - GFP tagged	RNA FISH	This study, performed by Meredith Wouters
S. cerevisiae	BY4741 *SCW10* - GFP tagged	RNA FISH	This study, performed by Meredith Wouters
S. cerevisiae	BY4741 *SED1* - GFP tagged	RNA FISH	This study, performed by Meredith Wouters
S. cerevisiae	BY4741 *TIM17* - GFP tagged	RNA FISH	Brown *et al.* 2018
S. cerevisiae	BY4741 *URA4* - GFP tagged	RNA FISH	Brown *et al.* 2018
S. cerevisiae	*spp1*Δ *ARG4* - GFP tagged	RNA FISH	This study, performed by Meredith Wouters
S. cerevisiae	*spp1*Δ *ATP4* - GFP tagged	RNA FISH	This study, performed by Meredith Wouters
S. cerevisiae	*spp1*Δ *BAT2* - GFP tagged	RNA FISH	This study, performed by Meredith Wouters
S. cerevisiae	*spp1*Δ *GCV3* - GFP tagged	RNA FISH	This study, performed by Meredith Wouters
S. cerevisiae	*spp1*Δ *GDH1* - GFP tagged	RNA FISH	This study, performed by Meredith Wouters
S. cerevisiae	*spp1*Δ *HMS2*-GFP tagged	RNA FISH	This study, performed by Meredith Wouters
S. cerevisiae	*spp1*Δ *HOM3* - GFP tagged	RNA FISH	This study, performed by Meredith Wouters
S. cerevisiae	*spp1*Δ *KAP104* - GFP tagged	RNA FISH	This study, performed by Meredith Wouters
S. cerevisiae	*spp1*Δ *SCW10* - GFP tagged	RNA FISH	This study, performed by Meredith Wouters
S. cerevisiae	*spp1*Δ *SED1* - GFP tagged	RNA FISH	This study, performed by Meredith Wouters
S. cerevisiae	*spp1*Δ *TIM17* - GFP tagged	RNA FISH	This study, performed by Meredith Wouters
S. cerevisiae	*spp1*Δ *URA4* - GFP tagged	RNA FISH	This study, performed by Meredith Wouters
S. cerevisiae	BY4741	NET-seq	Churchman & Weissman 2011
S. cerevisiae	*eaf3*Δ	NET-seq	Churchman & Weissman 2011
S. cerevisiae	*rco1*Δ	NET-seq	Churchman & Weissman 2011
S. cerevisiae	*set1*Δ	NET-seq	Churchman & Weissman 2011

Organism	Background	Experimental data	Source
S. cerevisiae	*set2*Δ	NET-seq	Churchman & Weissman 2011
S. cerevisiae	*paf1*Δ	NET-seq	Fischl *et al.* 2017
S. cerevisiae	*spt4*Δ	NET-seq	This study, performed by Ülkü Uzun
S. cerevisiae	*dst1*Δ	NET-seq	This study, performed by Ülkü Uzun
S. cerevisiae	*xrn1*Δ	NET-seq	This study, performed by Harry Fischl
S. cerevisiae	W303-1a	PRO-seq	Booth *et al.* 2016
S. cerevisiae	W303-1α *spt4*Δ	PRO-seq	Booth *et al.* 2016
S. cerevisiae	CEN.PK	NET-seq at 12 time-points in the YMC	This study, performed by Shidong Xi
S. cerevisiae	BY4741	NET-seq	This study, provided by François Robert
S. cerevisiae	Polymerase II CTD mutant, tyrosine_1 to Phenylalanine	NET-seq	Provided by François Robert (Collin *et al.* 2019)
H. sapiens	HeLa	mNET-seq, DMSO and DRB treated	This study, performed by Michael Tellier
H. sapiens	HeLa	RNA-seq	Mayer *et al.* 2015

1.10 Aims of research

Here we aim to answer a number of questions which feed into the idea of transcription as a fundamentally dynamic process, which is currently lacking finer detail regarding the transitions the polymerase makes and how these transitions are governed. We aim to determine how widespread observed phenomena are, across both the *S. cerevisiae* genome and then later across species in human cell lines. In Chapter 2 we ask how closely linked are the positions of nucleosomes at individual genes and the transcriptional state of these genes. It has been observed at a very small number of genes in the YMC that the nucleosomes around the promoter are dynamic at genes undergoing transient activation and repression, we wish to widen this study to a more genome-wide view. By looking at nascent and mature transcript levels on cells undergoing the shift from glucose to galactose, we utilize a system in which transcription is rapidly and transiently altered genome-wide. Coupling this data with nucleosome-positioning data via MNase-seq we determine the correlation between the position of nucleosomes around the promoter with activation and repression of transcription.

In Chapter 3, we wish to understand how factors which have been speculated to control transcription affect the dynamics of RNA synthesis and processing. In order to answer this, we develop a stochastic model of RNA production and synthesis and use this to analyse RNA FISH data at a number of loci. By finding the parameters of transcription that replicate the RNA FISH distributions across a population of cells, we determine the parameters of transcription initiation, elongation and RNA turnover. By further performing interventions involving altering the histone post-translational modifications and the levels of antisense transcription across genes, we determine how these regulatory factors affect the dynamics of RNA synthesis at these loci. In this Chapter we find that antisense transcription alters the rates of RNA synthesis and degradation in a manner similar to the changes brought about by high levels of histone acetylation, suggesting that any effects of antisense transcription on sense transcription dynamics may be through associated changes to the chromatin.

In Chapter 4 we develop bioinformatic tools to gain in-depth insights into the profiles of actively transcribing polymerases in *S. cerevisiae* and human cells. Previous studies aiming to unpick how polymerase proceeds through the genes on the DNA have relied on averaging across all genes in an organism, hiding the heterogeneity of the underlying profiles of polymerase which are seen across the *S. cerevisiae* genome (Churchman & Weissman 2011, Booth *et al.* 2016). We show that the profiles of transcriptionally-engaged polymerase across the genome are distinct not only between species, but across different genes within the same organism. By looking at how the distributions change upon intervention of the regulatory network, we highlight that not every gene is the same and different factors contribute to create different polymerase profiles across individual genes, most likely through different transcriptional dynamics.

Finally in Chapter 5 we develop a complete model of polymerase elongation, motivated by the bioinformatic work set out in Chapter 4. While in Chapter 4, we are able to qualitatively describe the profiles of polymerase across the genome, we wish to add quantitative understanding to these profiles in the context of the known processes undertaken by the RNA PolII complex. We first determine a minimal best-fit model which accurately captures the profiles of polymerase determined by NET-seq, parameterised at every gene

by the raw data. By fitting the model to NET-seq data in wild-type cells and following genetic intervention, we determine the effect of a number of regulatory factors which are either known to directly interact with the elongation complex, or have been speculated to have either direct or indirect effects on the process of transcription. Putting these results together, we are able to begin piecing together the roles of the many factors associated with the transcription process in setting up the kinetics of transcription.

Chapter 2

Gene regulation is dynamic upon change of media

The work in this chapter was carried out with Harry Fischl and Ronja Woloszczuk. NET-seq and RNA-seq experiments were performed by Harry and MNase-seq experiments were performed by Ronja. The bioinformatic work was done in close collaboration with Harry.

2.1 Overview

One drawback of many genome-wide studies is that they are performed in quasi-steady-state conditions, either following the deletion of a gene, the removal of a functional protein or after a lengthy treatment with an inhibitory chemical. Without knowledge of the dynamics of the underlying system, we are not able to distinguish potential sub-classes of behaviours that individual cells or even genes may undergo following either natural perturbations or external stimuli. In this chapter we aim to determine the relationship between the dynamics of gene expression and the chromatin landscape at genes which undergo changes in their transcriptional activity.

By using a dynamic system whereby we are able to distinguish both which genes are undergoing changes in expression as well as when these changes occur, we move beyond the problems found with static steady-state data. Here, we ask how nucleosome positions and gene expression are linked upon a change of carbon source from glucose to galactose. The changes in metabolism required upon this shift of media results in both local expression

changes at the *GAL* genes, directly responsible for the metabolism of galactose (Venturelli *et al.* 2012), but we also observe genome-wide changes in expression. This results in an ideal system for us to unpick the interplay between chromatin and gene expression.

2.2 Deep sequencing techniques allow for analysis of transcript dynamics

High-throughput sequencing methods have allowed the move from analysis of single genes to a more global genome-wide view of the distributions of RNA, nucleosomes and regulatory proteins, such as transcription factors, bound to the DNA. While on their own each experiment offers a static view of the current state of the cell, by performing these experiments through a time-course following perturbation we can obtain a dynamic view of gene regulation in cells. In this Chapter, we focus on the link between the positions of nucleosomes and the expression of transcription units.

It has been suggested that the links between the placement of nucleosomes and the activity of genes may not be a straight-forward one. In Murray *et al.* 2015 it was proposed that the placement of the first nucleosome to be transcribed through may dictate the expression level of a gene through the availability of transcription factor binding sites or the TSS. By either placing a nucleosome downstream of the binding site of an activating transcription factor or occluding the binding site of a repressive transcription factor, the placement of this nucleosome may activate transcription, or repress it through the opposing action. Alternatively, the placements of the nucleosomes may be a consequence of transcriptional activity, with the activity of the transcription factors and polymerase dictating the locations of the nucleosomes (Knight *et al.* 2014, Nocetti & Whitehouse 2016). Furthermore, the compaction of the nucleosomes may determine the density of the chromatin environment, creating feedback-loops of transcriptional activity by forming transcriptional "hubs" of activity (Hnisz *et al.* 2017).

We find that transcription of sub-groups of genes are coordinated upon change of carbon source and these sub-groups of genes are approximately grouped into their role in the metabolic state of the yeast cell (Tu *et al.* 2005) and that these changes in transcription

are mirrored later in time at the levels of transcripts. Furthermore, we find that contemporaneous with a change in expression, the nucleosome found immediately downstream of the TSS is dynamic, moving towards the TSS when genes are repressed and downstream upon activation.

The use of NET-seq, RNA-seq and MNase-seq (Churchman & Weissman 2011, Kent & Mellor 1995, Knight *et al.* 2014) in cells undergoing this change in carbon source uncovers genome-wide changes in transcription, with the cells rapidly changing their chromatin landscape and expression within minutes of an external cue.

2.2.1 The Yeast Metabolic Cycle demonstrates cycling sub-classes of genes

S. cerevisiae cells have been found to follow cycling metabolic activity, in particular a population of yeast cells in a chemostat cycle in a synchronised manner, reflected in the amount of dissolved oxygen over time (Tu *et al.* 2005). Furthermore, a subset of approximately 57% of genes were found to cycle with the phase of the metabolic cycle observed. These genes were found to express in the so-called oxidative, reductive charging and reductive building stages of the yeast metabolic cycle (Figure 2.1), hence ascribing the ontologies to these genes. Nocetti & Whitehouse 2016 found that in a synchronised population of cells, a small number of genes exhibited shifts in the location of the plus one nucleosome, found abutting the transcription start site, upon gene activation. Here, we describe a system where a large number of genes exhibit changes in gene expression dynamics and ask what behaviours the nucleosomes follow. When cells undergo the change of media from glucose to galactose, we show that a large number of genes exhibit rapid activation and repression and use this as a system to study the interaction between nucleosome positions and expression at fine detail. Using this system we are able to show that those behaviours exhibited at a small number of genes in the YMC are in fact widely seen across the whole genome when gene expression dynamic changes are fast.

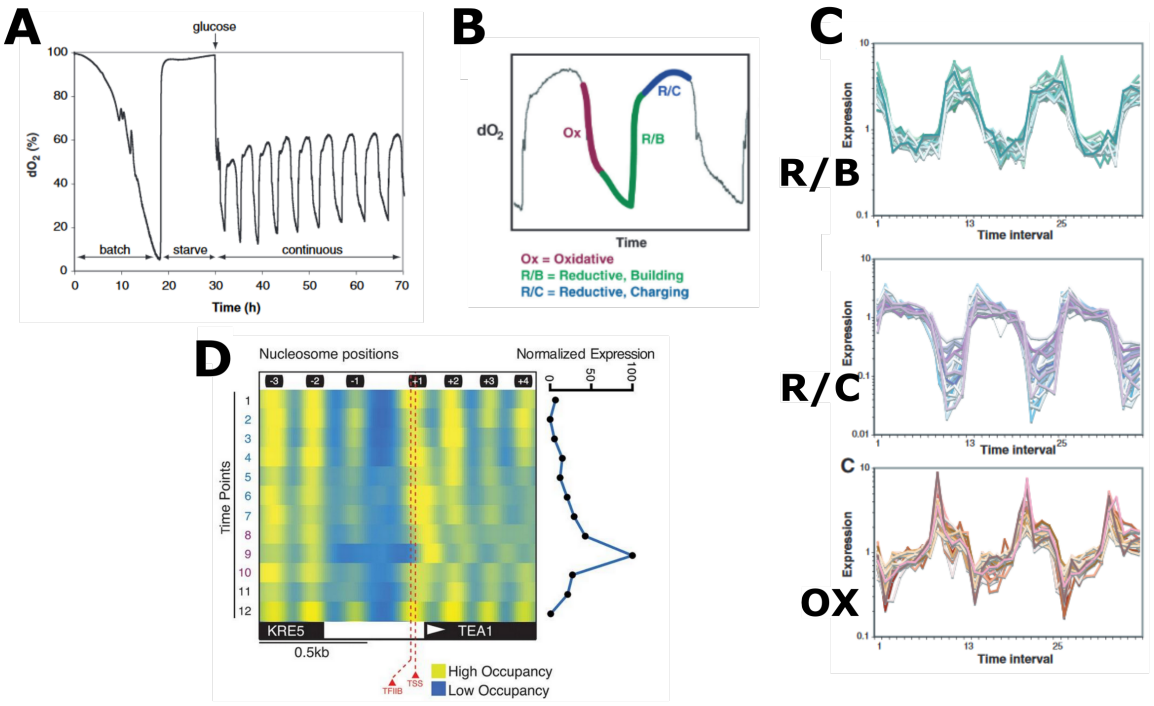


Figure 2.1: A summary of observations from the yeast metabolic cycle. Under glucose starvation conditions, yeast cells synchronise their metabolism as observed by dissolved oxygen concentration (A and B). Transcripts exhibit cycling behaviour as found by RNA sequencing. Shown are the top 10 genes that fall into the Reductive Building (RB), Reductive Charging (RC) and Oxidative (OX) phases (C), taken from Tu *et al.* 2005. A sample gene from Nocetti & Whitehouse 2016 shows the plus one nucleosome moves contemporaneously with transcript levels (D).

2.2.2 MNase-digestion followed by deep sequencing shows locations of nucleosomes

To analyse the locations of nucleosomes across the genome, we turned to MNase-seq, where micrococcal-nuclease (MNase) is used to digest exposed fragments of DNA, leaving those regions which are inaccessible to the digestion enzyme (Figure 2.2, Kent & Mellor 1995, Knight *et al.* 2014). DNA is cross-linked to any bound protein, then MNase digests exposed fragments of DNA not protected by any factors. By further selecting for DNA fragments of a size comparable to a nucleosome (∼150bp), the fragments correspond to DNA wrapped around a nucleosome, rather than smaller regions likely to correspond to proteins which bind to the DNA such as transcription factors. Aligning these fragments to the *S. cerevisiae* genome creates a series of peaks and troughs of reads, corresponding to a probability distribution of the locations of nucleosomes. Sharper peaks correspond to less "fuzzy" nucleosomes, with a higher probability of being in a single location across a pop-

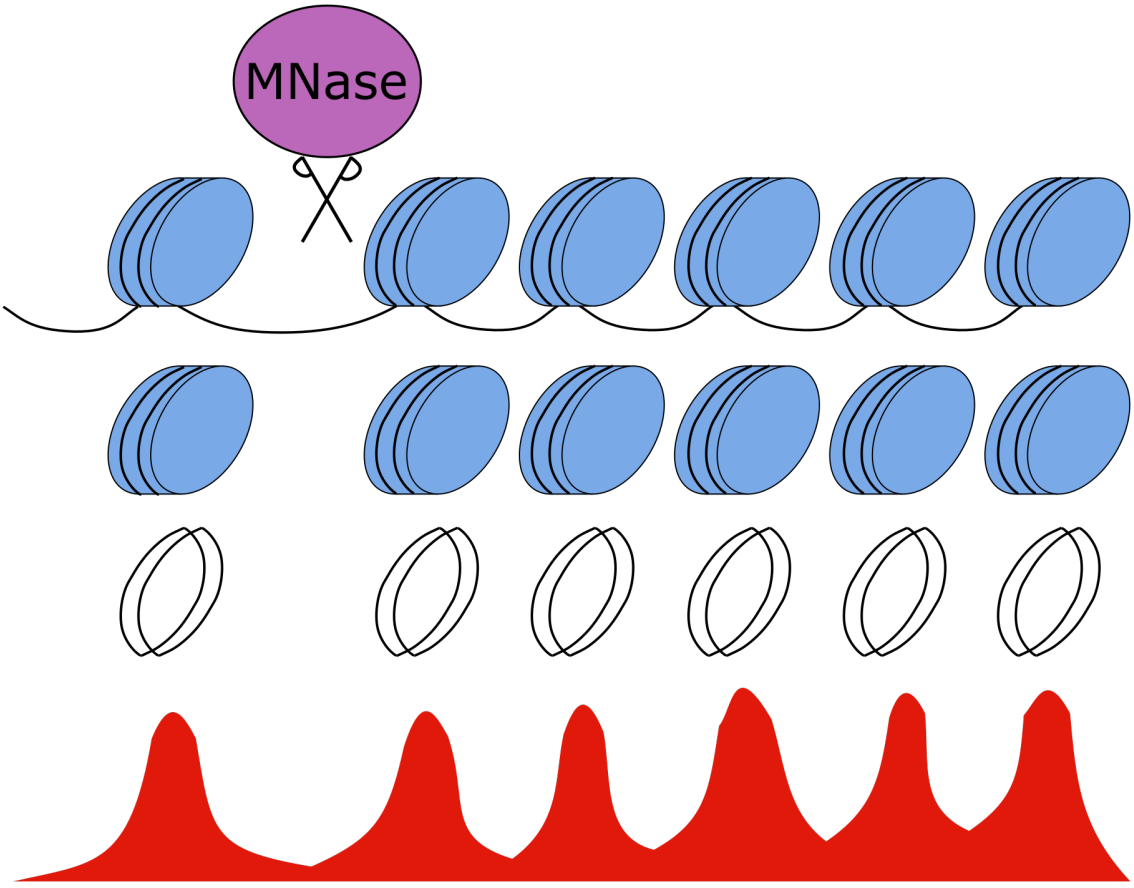


Figure 2.2: MNase-seq work-flow (top to bottom), exposed DNA is digested by MNase, leaving cross-linked DNA wrapped around nucleosomes. Nucleosomes and other proteins are then degraded, with remaining DNA fragments taken forward for sequencing. Finally, DNA sequences are aligned to the genome, giving a probability distribution of the locations of nucleosomes across the genome.

ulation of cells, whereas the opposite suggests less well-defined positions of nucleosomes.

To further find the locations of individual nucleosomes, the aligned sequences were parsed through the peak-calling software DANPOS (Chen *et al.* 2013), with minimum inter-peak distance set to 146bp. This gives an estimate for the centre of each nucleosome from the MNase-seq data. We ran DANPOS for three datasets: *S. cerevisiae* cells grown in glucose-containing media, cells grown in glucose then transferred to galactose-containing media for 15 minutes and cells transferred to galactose-containing media for 60 minutes. We thus were able to locate each nucleosome at loci across the yeast genome throughout the glucose-galactose shift.

A full pipeline of the MNase-seq alignment, processing and analysis can be found

at: `https://github.com/tbrown91/Scer_MNasePipeline`. Briefly, single-end sequencing after MNase-digestion were aligned to *S. cerevisiae* genome release SacCer3 using Bowtie2 version 2.2.6:

```
bowtie2 -x Genome_build -U seq.fastq -S aligned_seq.sam
```

Danpos version 2.2.2 was then run on sorted bam files with minimum inter-nucleosomal distance of 146bp:

```
python ./danpos-2.2.2/danpos.py dpos bam_files/ -o DANPOS_OUTPUT/ -jd 146 -a 1
```

2.2.3 Nascent and mature RNA sequencing give high-resolution views of gene expression dynamics

To analyse any gene expression changes upon change of carbon source media, we performed nascent RNA sequencing (NET-seq), sequencing RNA found in the active site of polymerase and total RNA sequencing (RNA-seq) for *S. cerevisiae* cells grown in glucose-containing media, and after 5, 15 and 60 minutes in galactose-containing media. We use NET-seq for the analysis of changes in transcription, as this gives direct measurements of transcriptional activity and changes in these levels (Churchman & Weissman 2011). Unlike with RNA-seq, using NET-seq as a measurement of transcription we are not reliant on the assumption that nuclear or cytoplasm degradation will not be affected by changing the media from glucose to galactose.

By using NET-seq and RNA-seq as measurements of gene expression, we are able to pin-point the timings of gene activation and repression from the measurements of nascent transcription and furthermore identify the potential time delays to seeing these changes on levels of mature RNA via RNA-seq. We are also able to see the changes in expression levels of ncRNA using NET-seq, capturing the total levels of transcription at these loci before they are targeted for early degradation. Using these techniques in tandem with the locations of the nucleosomes allows us to gain a complete picture of the dynamics of gene expression and its interplay with the chromatin landscape.

Gene expression changes globally on change of media

Across the yeast genome, a large number of genes exhibited at least a 2-fold change in relative nascent transcription levels and had a defined plus one nucleosome in the first 100bp after the TSS (∼720 genes). These genes were taken forward for further analysis and clustered into 4 groups (Figs 2.4-2.7) where clusters 1 and 2 showed increase in relative nascent transcript levels (Figs 2.4B & 2.5B). Clusters 3 and 4 exhibited immediate drops in relative numbers of nascent transcripts (Figs 2.6B & 2.7B).

In all clusters, we found that changes in initial levels of nascent transcript resulted in a delayed change in the levels of relative mature RNA (Figs 2.4B - 2.7B). This suggests that upon change of media, immediate changes to transcription are not buffered at the level of transcripts and indeed, these changes persist after 60 minutes. We then asked what is occurring to the chromatin landscape at these genes that exhibit changes in gene expression. The changes to gene expression upon the shift to galactose media also closely fall into cells moving out of an oxidative state into a reductive/charging state, as defined by the YMC, reflecting an alternative metabolic state of the cells (Tu *et al.* 2005, see Figures 7.1 - 7.3).

2.3 Plus one nucleosome position moves with gene expression activity

Performing MNase-digestion followed by sequencing allows us to determine the locations of nucleosomes across the *S. cerevisiae* genome. We are particularly interested in those nucleosomes surrounding the transcription start site and how these nucleosome positions are determined in relation to gene activation and repression. It has been previously reported at a small number of genes that the plus one nucleosome, situated immediately downstream of the TSS, is dynamic upon gene activation in the yeast metabolic cycle (Nocetti & Whitehouse 2016). Here, we find the centres of all nucleosomes in yeast cells grown in glucose media, then 15 and 60 minutes following change of media to one containing galactose as carbon source. The *GAL*-system of genes responsible for metabolism of galactose is well-studied as is the network of interactions upon induction (Venturelli

et al. 2012). Here, we wish to use this system to identify sets of genes which undergo expression changes and ask whether these changes in expression are linked to chromatin.

First, taking a genome-wide view of nucleosome positions around gene TSSs reveals a much more relaxed chromatin environment in galactose-containing media than glucose conditions, as determined by a wider NDR, which is coincident with gene promoters (Figure 2.3). While there are not consistent changes in the positions of nucleosomes within gene bodies (+1, +2, +3 nucleosomes, Figure 2.3B), there is a clear shift upstream of the -1 nucleosome found upstream of the TSS (Figure 2.3B). Furthermore, defining the nucleosome-depleted region (NDR) as the distance between the -1 and +1 nucleosome at each gene, we see larger NDRs detected at promoters in galactose-containing media (Figure 2.3C). While previous studies have associated open promoter chromatin with higher gene activity (Zaugg & Luscombe 2012), we wished to test the strength of this association in further detail.

In order to determine the relationship between nucleosome positioning and gene expression, we took nascent NET-seq and steady-state RNA-seq experiments performed in glucose-containing media and after 5, 15 and 60 minutes in galactose-containing media. We grouped genes via soft-clustering into 4 clusters which resulted in groups of genes showing either a general increase in nascent transcription levels after a shift to galactose media (Figures 2.4 & 2.5) or those with a general decrease in nascent transcription levels (Figures 2.6 & 2.7).

We took forward those genes which had a detected nucleosome in a window from the annotated TSS to 100bp downstream of the TSS. Relative plots of the level of MNase at each gene ordered by the distance from the detected +1 nucleosome to the annotated TSS are shown in Figures 2.3A–2.7A. The +1 and +2 are visibly well-positioned at most genes as shown by these heatmaps, and the nucleosome-depleted region defined as the distance from the -1 nucleosome to the +1 nucleosome is seen widening around the TSS from glucose to galactose.

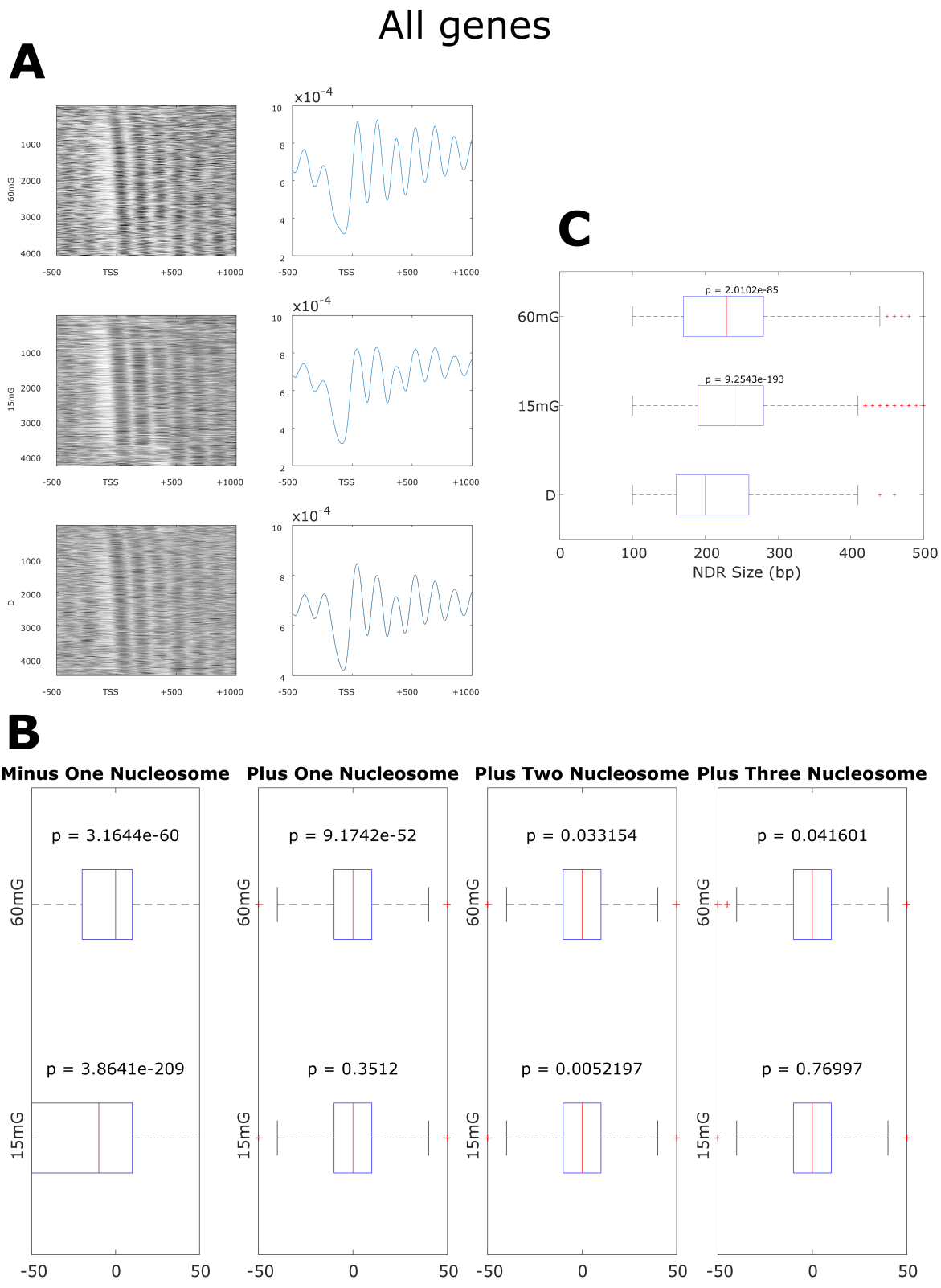


Figure 2.3: Positions of -1, +1, +2 and +3 nucleosomes at all genes upon change of media from glucose (D) to galactose (G). Shown are the positions of the nucleosomes at 15 and 60 minutes in galactose relative to glucose. A positive shift represents a 5' → 3' shift and vice-versa. RNA sequencing experiments performed by Harry Fischl and MNase-seq by Ronja Woloszczuk.

2.3.1 Nucleosomes move into the gene body at activated genes

We found that while the upstream shift of the -1 nucleosome is consistent between clusters of genes independent of changes in expression level, the position of the +1 nucleosome varies with gene expression changes. In those clusters with increased gene expression (Figs. 2.4 & 2.5) we observed a downstream shift in the +1 nucleosome at 15 minutes in galactose, which was somewhat diminished by 60 minutes (Figs. 2.4C & 2.5C). We also observed a slight downstream shift in the +2 nucleosomes at these genes, concurrent with the idea of nucleosomes forming a strict tiling pattern at expressed genes, however by the +3 nucleosome we observed no significant changes in the positions of nucleosomes this far into the gene body.

2.3.2 Nucleosomes move towards the promoter at repressed genes

Taking those genes that showed a decrease in nascent transcription activity revealsed that, consistent with those showing increased transcriptional activity, there was a general opening of the nucleosome-depleted region around the TSS (Figs 2.6D & 2.7D), dictated again by a large upstream shift in the -1 nucleosome (Figs 2.6C & 2.7C). However, independent of the apparent opening of the promoter chromatin, these repressed genes exhibited an upstream shift in the location of the +1 and +2 nucleosomes, bringing them closer to the TSS. Similarly to those genes with increased activation, the shifts in nucleosomes were diminished by 60 minutes compared to 15 minutes, suggesting highly dynamic nucleosomes moving on this shift in carbon source media. Again, the fact that the +1 and +2 nucleosomes track each other suggest that even though these nucleosomes may be dynamic in this system, the even tiling of the nucleosomal array is preserved at all genes.

2.4 Discussion

The work presented in this chapter linking nucleosomal dynamics with gene transcriptional activity highlights the important link between the transcription complex and the chromatin at genes. It is tempting to immediately jump to the conclusion that the +1 nucleosome defines the expression status of a gene, especially if we believe this nucleosome

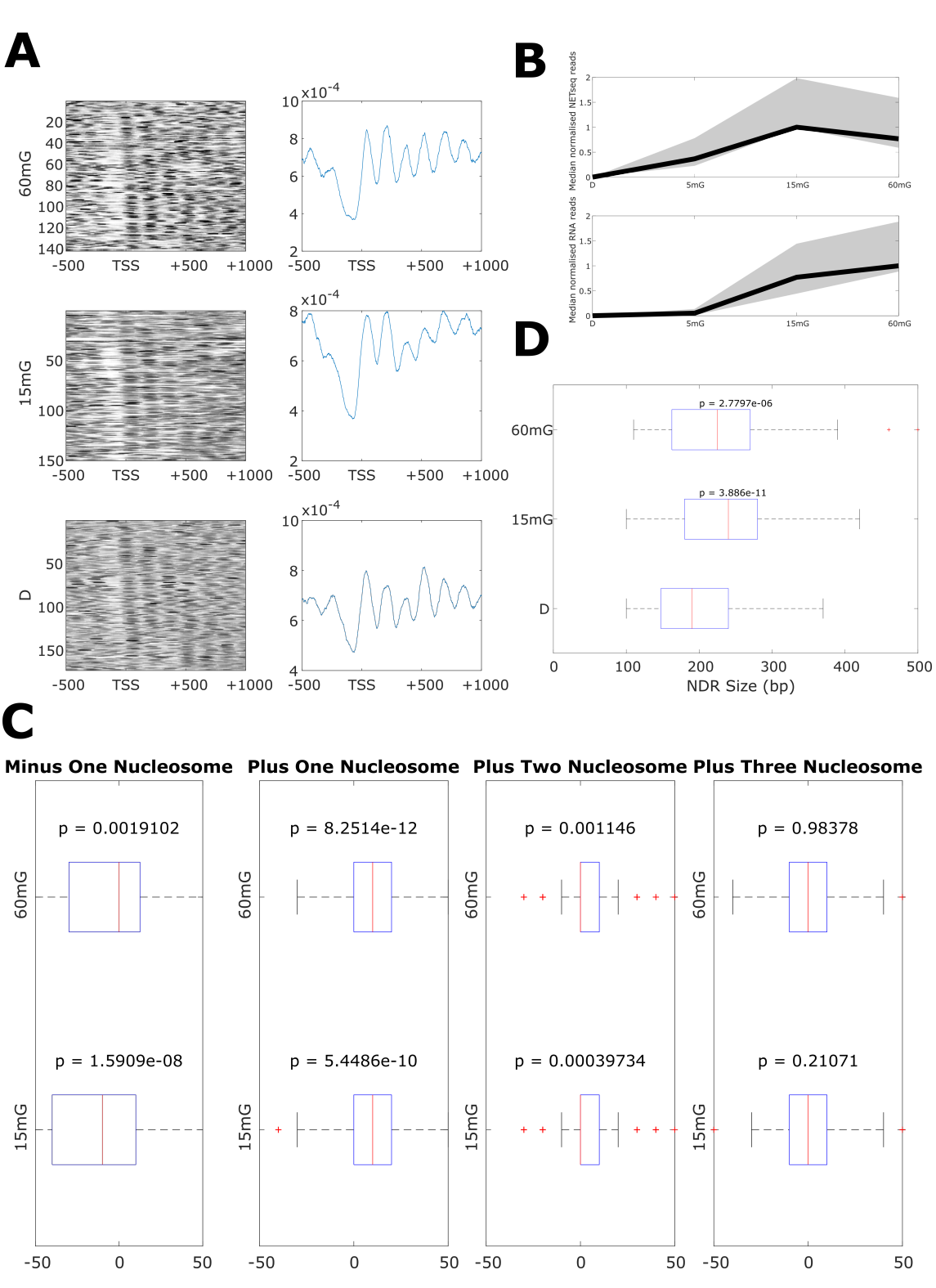


Figure 2.4: Positions of -1, +1, +2 and +3 nucleosomes at genes in cluster 1 upon change of media from glucose to galactose. Shown are the positions of the nucleosomes at 15 and 60 minutes in galactose relative to glucose. A positive shift represents a 5' $\rightarrow$ 3' shift and vice-versa. RNA sequencing experiments performed by Harry Fischl and MNase-seq by Ronja Woloszczuk.

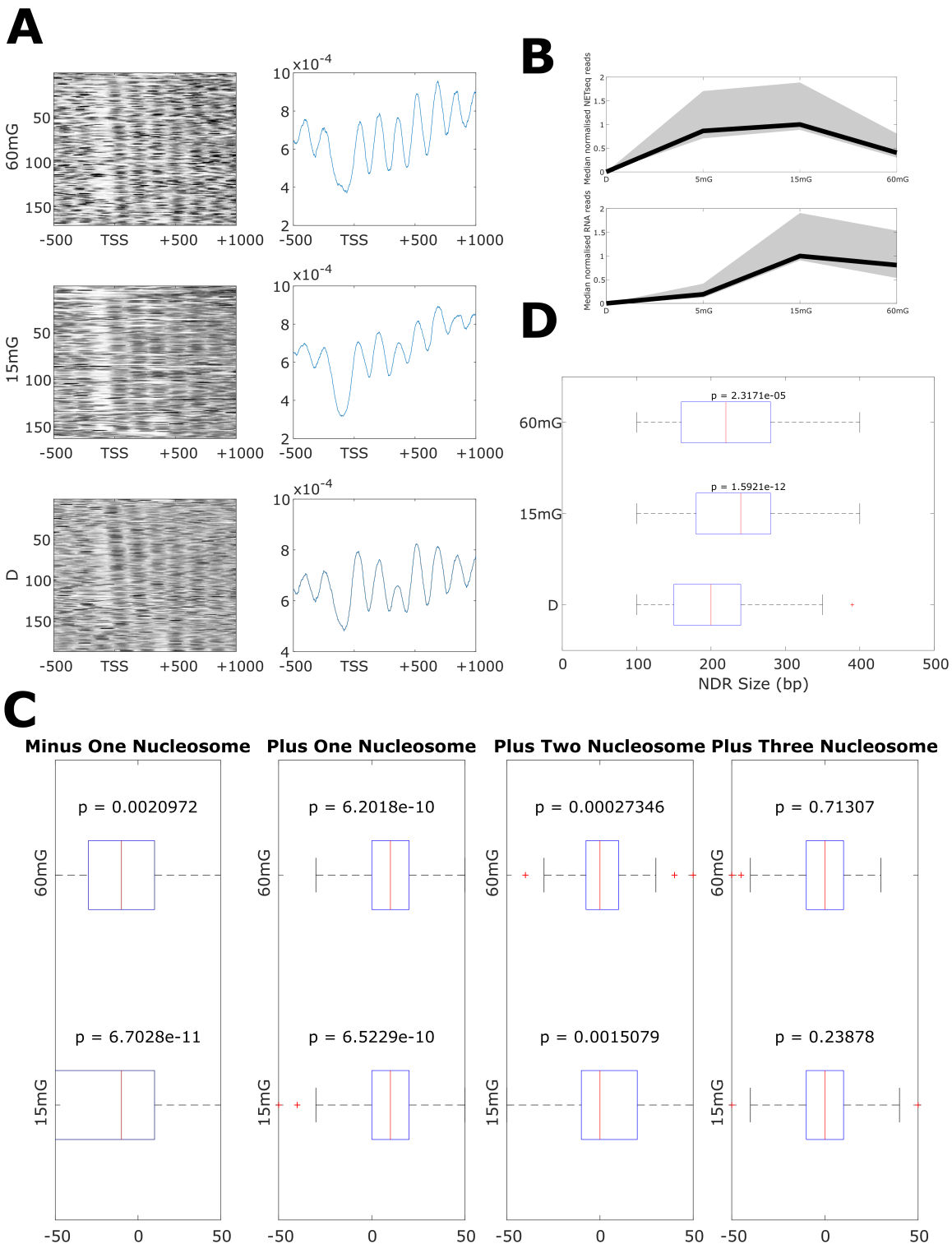


Figure 2.5: Positions of -1, +1, +2 and +3 nucleosomes at genes in cluster 2 upon change of media from glucose to galactose. Shown are the positions of the nucleosomes at 15 and 60 minutes in galactose relative to glucose. A positive shift represents a 5' $\rightarrow$ 3' shift and vice-versa. RNA sequencing experiments performed by Harry Fischl and MNase-seq by Ronja Woloszczuk.

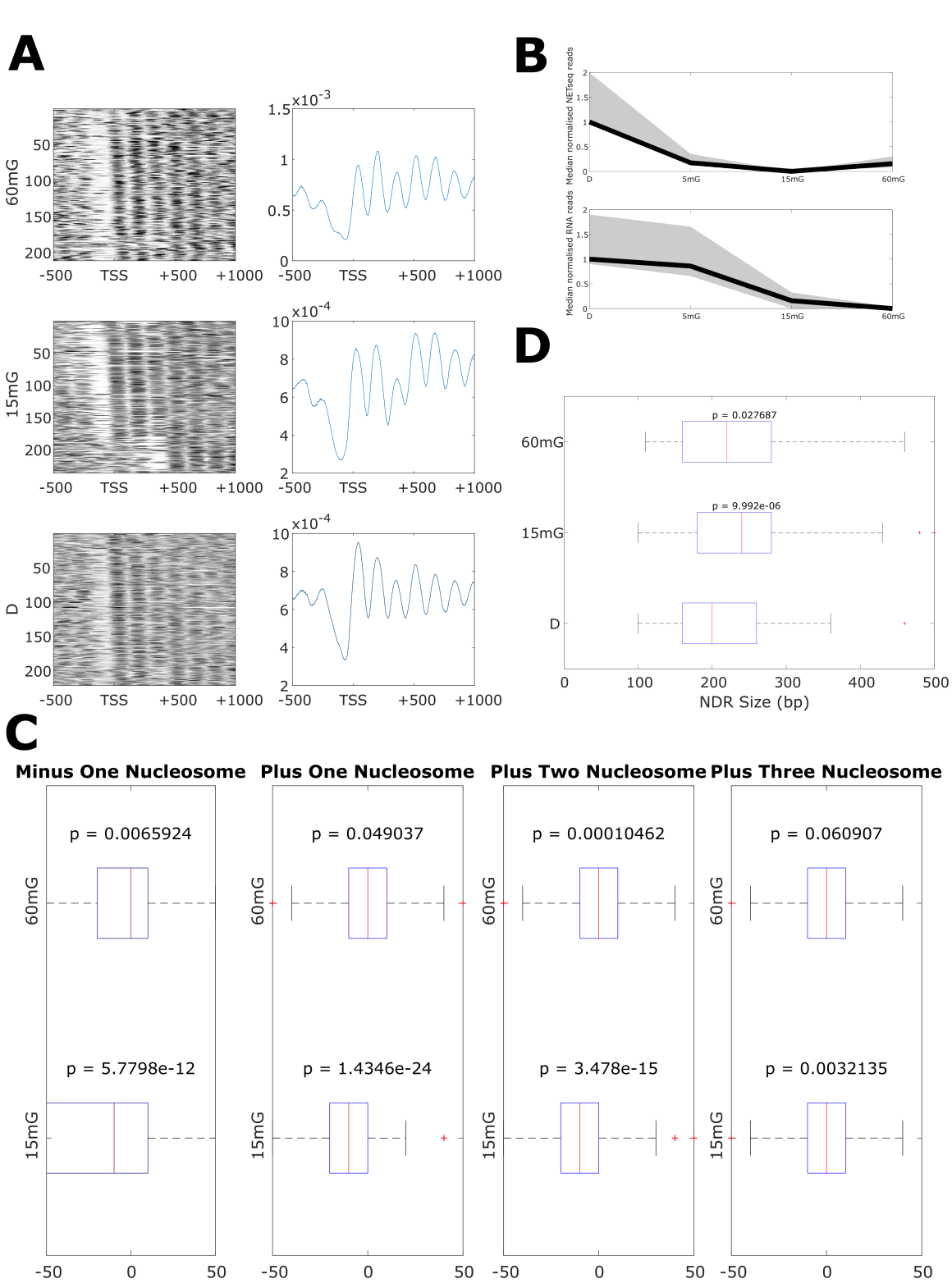


Figure 2.6: Positions of -1, +1, +2 and +3 nucleosomes at genes in cluster 3 upon change of media from glucose to galactose. Shown are the positions of the nucleosomes at 15 and 60 minutes in galactose relative to glucose. A positive shift represents a 5' $\rightarrow$ 3' shift and vice-versa. RNA sequencing experiments performed by Harry Fischl and MNase-seq by Ronja Woloszczuk.

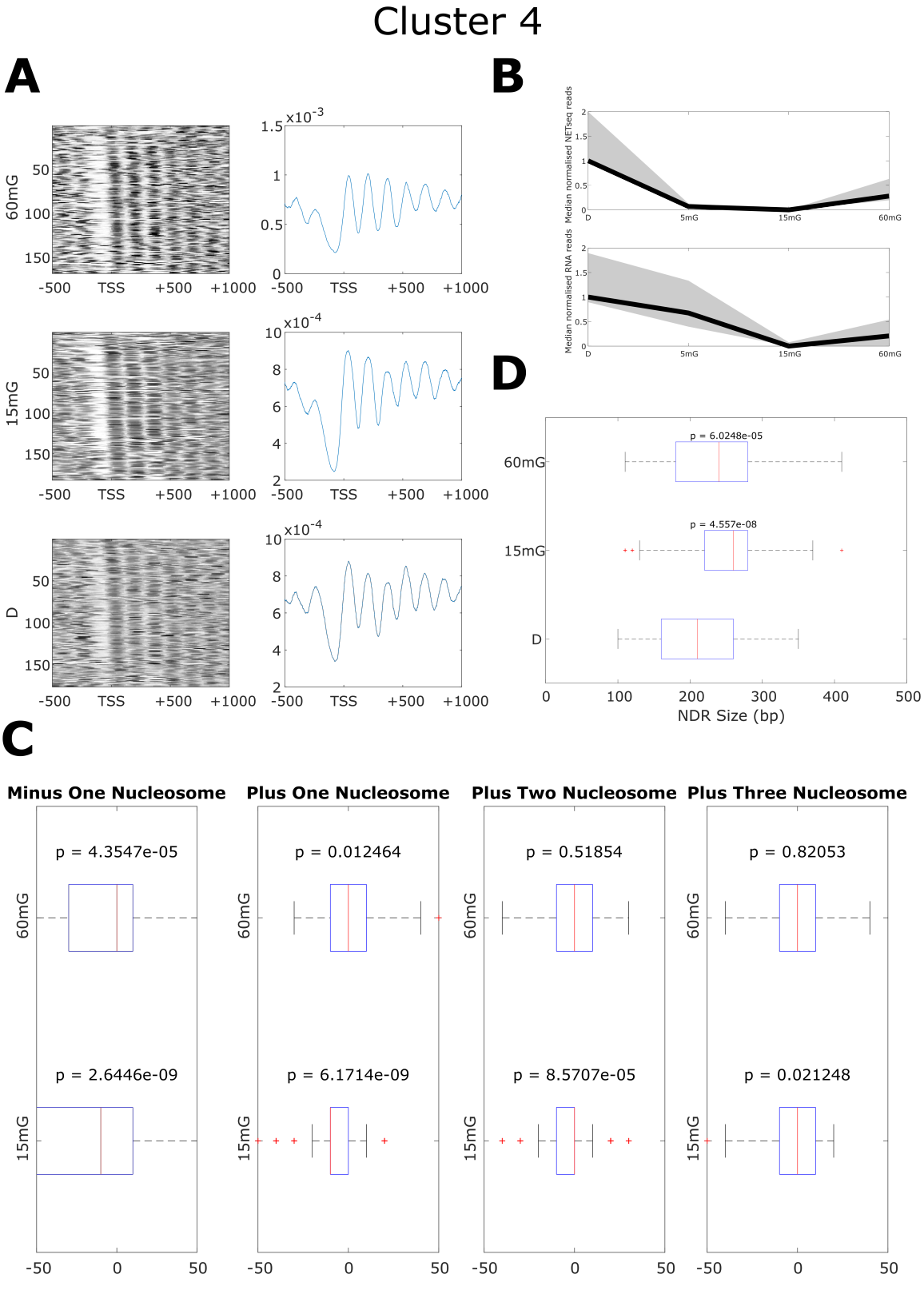


Figure 2.7: Positions of -1, +1, +2 and +3 nucleosomes at genes in cluster 4 upon change of media from glucose to galactose. Shown are the positions of the nucleosomes at 15 and 60 minutes in galactose relative to glucose. A positive shift represents a 5' → 3' shift and vice-versa. RNA sequencing experiments performed by Harry Fischl and MNase-seq by Ronja Woloszczuk.

occludes the TSS and key binding sites for polymerase and transcription factors as has been suggested previously (Knight *et al.* 2014, Murray *et al.* 2015).

We cannot, however, rule out the possibility that the positioning of nucleosomes is defined by transcriptional activity. In the absence of elongating polymerase, it has been shown that well-defined nucleosomal arrays are not reconstituted (Hughes *et al.* 2012). Therefore, we may conclude that transcription elongation is an essential process for the establishment of the typical well-spaced arrays of nucleosomes and the exact locations of the arrays are dependent on the expression levels at each gene. The presence of activating transcription factors at the promoter of those genes with increased expression is likely to create greater torsional stress on the DNA wrapped around chromatin. Indeed, this would easily explain the pushing downstream of the +1 and +2 nucleosomes at those activated genes (Figs. 2.4 & 2.5). Equally, we could imagine those repressed genes losing these transcription factors from their promoter regions, relaxing the DNA and "springing" the nucleosomes back to a more neutral position closer to the promoter.

The same could also be true of the polymerase itself. It may be the case that the presence or the activity of an elongation polymerase binding to the promoter region and initiating transcription could push the nucleosomes from the NDR into the gene body. Similarly, the loss of the polymerase from repressed genes could result in the +1 nucleosome moving upstream to occupy the NDR.

However, the true underlying order cannot be that simple, as not all transcription factors activate transcription. A number of genes are regulated through the activity of repressive transcription factors binding promoter regions to shut off transcription (Courey & Jia 2001, Gaston & Jayaraman 2003) or entire repressive complexes binding the promoter region, like the Polycomb Repressive Complex proteins in higher eukaryotes (Entrevan *et al.* 2016). Furthermore, the non-uniform distributions of the locations of transcription factor binding sites relative to the TSS adds more complexity. These factors together perhaps rule out interplay between the placement of transcription factors and nucleosomes and the true order may either be the downstream act of transcription setting up the nucleosome placement, or the nucleosome environment signaling to other factors the location of a gene ready for active transcription.

Clearly the true picture regarding the link between chromatin and gene expression is complicated and goes well beyond simply the placement of nucleosomes relative to the TSS. The histones are known to undergo a great deal of post-translational modification as different factors are brought to the chromatin (Mellor 2006) and both histone modifications and transcripts oscillate at a large proportion of genes in cells synchronised in the YMC (Kuang *et al.* 2014). In order to truly probe the order of events one would need to add further experiments to the set-up described above. By looking at the levels of known transcription factors at the promoters of genes analysed here, one could determine the importance of these factors in the expression levels of genes and how these link to the positioning of the nucleosomal array. By further measuring the levels of individual histone modifications at the nucleosomes of the repressed and activated genes, it would be possible to identify the key regulatory factors of the chromatin, beyond the positions of the nucleosomes.

Finally, one major drawback of the work carried out here is that the timings of the measurements taken do not have the resolution to pinpoint the cascade of events and whether expression follows changes in the chromatin or vice-versa. In order to answer this question, either experiments will have to be performed at finer resolution than 15 minutes after change in media, or the chromatin must be altered in another way to see if any direct effect results on gene expression. Taking observations on a finer time-scale would allow us to determine if it is the case that a lot of factors come with the polymerase, operating with a scanning and searching model, searching for open promoters with the right topological setup which must exist before transcription may occur.

The work presented here highlights the potentially crucial connection between the nucleosome arrays and correct expression of genes. There are a number of further speculated factors that may be involved in not only the overall expression, but the fine-tuning of the response to external cues governing transcription. We wish to further investigate how factors such as elongation factors, histone modifications and different forms of transcription affect the expression of genes and the dynamics of the RNA PolII complex.

Chapter 3

RNA production dynamics are modulated through histone acetylation

The experimental work in this Chapter was performed by Françoise Howe and Meredith Wouters. All experiments performed on the *GAL1* locus were performed by Françoise, the FISH data from wild-type and *spp1*Δ cells were performed by Meredith. The majority of the work in this chapter contributed towards the publication:

Brown, T. *et al.* Antisense transcription-dependent chromatin signature modulates sense transcript dynamics. *Molecular Systems Biology* **14,** 1–21 (2018).

Simulation scripts for this chapter are available at: `https://github.com/tbrown91/cuda_paramSearches`.

3.1 Transcription occurs through a complex environment

The conversion of DNA to RNA by RNA polymerase is pervasive in all genomes (Mellor *et al.* 2016), occurring on both strands of the DNA at protein-coding and non-coding templates. As such, this creates a picture of a crowded transcriptional environment, where RNA polymerase must transcribe through a DNA template found wrapped around nu-

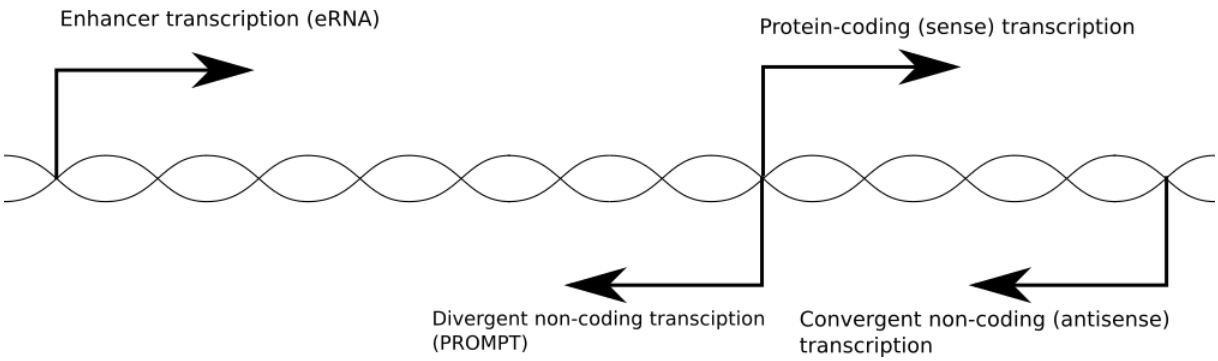


Figure 3.1: Types of transcription units in eukaryotic organisms. Two types of non-coding transcription are described relative to a protein-coding transcription unit: antisense transcription, originating downstream of the sense TSS and read in the reverse orientation, and divergent-antisense transcription units, originating in a sense promoter and read in the reverse orientation. Further non-coding transcription units exist such as enhancers, found in either inter- or intragenic regions.

cleosomes, arranged in 3-dimensional higher-order structures and with other regulatory factors associated with the DNA and chromatin. The relatively recent discovery of non-coding and antisense transcription in eukaryotic organisms has raised questions as to the purpose of these transcription units. Indeed the extent to which non-coding transcription is present suggests a role beyond merely background noise or faulty transcription initiation. The main transcription units discussed here are antisense transcription units, defined as non-coding transcription read in the reverse orientation to a coding transcription unit (See Figure 3.1).

The introduction of techniques tracking strand-specific nascent transcription such as PROseq (Kwak *et al.* 2013, Booth *et al.* 2016), GROseq (Core *et al.* 2008) or (m)NETseq (Churchman & Weissman 2011, Nojima *et al.* 2015) have highlighted the degree to which the genome is extensively transcribed outside of genic regions. Of particular focus in this chapter is the role of antisense transcription, a form of non-coding transcription originating within genes and reading in the reverse orientation. This form of transcription creates an environment where polymerases travel in both directions through a genic template, sharing the same chromatin landscape.

The identification of antisense transcription by population-averaged techniques such as northern blotting or RNA sequencing methods hides a more complex story regarding the interaction between the two transcription events. Using single-cell techniques such as RNA FISH or live-cell imaging of sense and antisense transcripts has identified locus-

specific control regarding the co-expression of these pairs. Castelnuovo *et al.* 2013 observed via RNA FISH at *PHO84* that in *S. cerevisiae*, cells were either found to be expressing the *PHO84* sense transcript or antisense transcript, but rarely showed co-expression of the two, with the same observation made at *HMS2* in Nguyen *et al.* 2014. Lenstra *et al.* 2015 however used live-cell imaging at *GAL10* and observed co-expression of the sense-antisense pairs as well as exclusive transcription of either individual transcript.

This complex transcriptional landscape creates a number of possible scenarios whereby polymerases may be firing convergently on opposite strands in rare scenarios. However it appears that at a number of genes, this scenario may not arise, with few head-on collisions of polymerases as described in Hobson *et al.* 2012. Given the pervasiveness of non-coding and antisense transcription in eukaryotic genomes, we need instead to look beyond this simplified idea of polymerase kinetics to unpick its role in gene regulation.

3.1.1 Transcription units are associated with unique transcription environments

Each of the transcription units described above are associated with distinct transcription complexes and/or chromatin environments. Active transcription of protein coding genes is associated with a polymerase complex with phosphorylated serine residues on the C-terminal domain of PolII as well as tri-methylation of lysine 4 on histone 3 (H3K4me3) and other active transcription marks (Howe *et al.* 2017, Santos-Rosa *et al.* 2002). On the contrary, high levels of antisense transcription are associated with a chromatin architecture with increased acetylation of histone 3 and decreased methylation of histone 3 or ubiquitination of histone 2b (Murray *et al.* 2015, Brown *et al.* 2018). The differences in the construction of sense and antisense polymerase complexes also differ, with antisense transcription polymerases showing no Serine-2 phosphorylation on their CTD (Murray *et al.* 2015) and alternative elongation factors, such as Paf1, have been shown to bind preferentially to sense-transcribing RNA PolII over antisense (Fischl *et al.* 2017). The same chromatin architecture has also been found associated with those divergent antisense transcription units found upstream of sense transcripts (Marquardt *et al.* 2014). Activity of enhancer transcription units has recently been shown to correlate with the ra-

tio of H3K4me1 to H3K4me3, again creating a transcriptional environment distinct from that at active protein-coding genes.

Clearly there is not one form of transcription, with transcription of different types of transcription units creating distinct environments for polymerase to progress through. However, given the crowded nature of the genome, it is not yet clear what role the transcription of non-coding transcription units plays, if any, in the regulation of protein-coding genes and in particular on the dynamics of gene expression.

Here we aim to understand the role of convergent antisense transcription, on the expression dynamics of the corresponding sense gene, which transcribes through the same chromatin environment in the opposing direction and how this role might be affected by other factors. Furthermore, we then aim to uncover what effect the placement of the mark of active transcription, H3K4me3, has on the dynamics of sense expression.

By combining mathematical modelling with quantitative experimental data we offer quantitative analysis of the roles of antisense transcription, histone modifications and the interplay between the two in controlling the dynamics of gene expression.

3.2 Modelling RNA FISH data reveals distinct transcription dynamics

3.2.1 RNA FISH

In this chapter we wish to determine how the dynamics of gene expression vary across the genome and determine which regulatory factors govern these changes. To model the kinetics of RNA production at high resolution, we require a technique that tracks single RNA molecules in single cells, therefore we turned to single molecule RNA FISH (Trcek *et al.* 2012). In RNA FISH, cells are fixed and flooded with single-stranded DNA fluorescent probes complementary to the RNA sequence of interest. Imaging the cells allows for quantification of the number of RNA molecules of our gene of interest in the cell at a fixed time. By further staining the DNA in the cell with DAPI (4',6-diamidino-2-phenylindole), we can localise RNA to either the nucleus or cytoplasm. By counting the number of RNA molecules in the nucleus and cytoplasm of hundreds of fixed cells, we

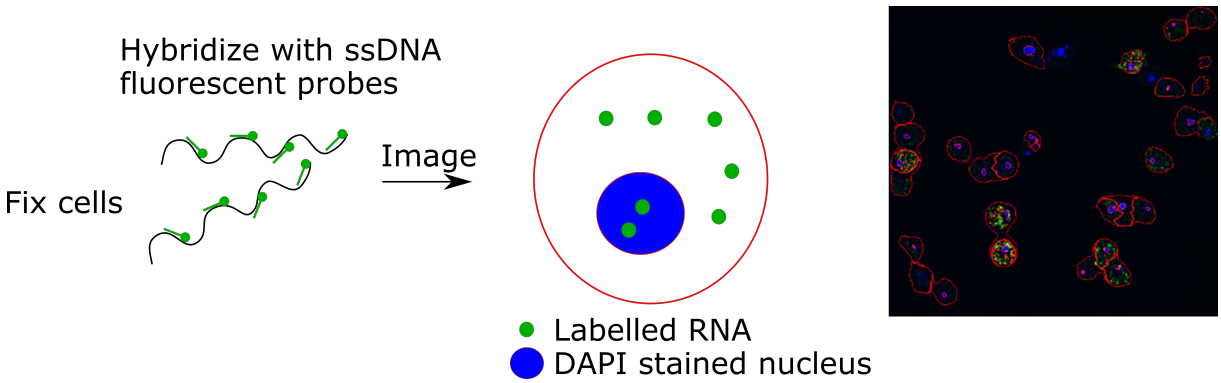


Figure 3.2: Single molecule RNA FISH labels the RNA sequence of interest by flooding the cells with fluorescent single-stranded DNA probes. Further staining of the nucleus with DAPI allows for the quantification and localisation of our RNA of interest to the nucleus or cytoplasm. FISH images were processed by Andrew Angel.

obtain a snap-shot of the distribution of nuclear and cytoplasmic RNA across a population of yeast cells (Figure 3.2).

3.2.2 Promoter dynamics

The nature of transcription initiation has come under increased scrutiny recently with the advent of new techniques analysing production of both nascent and stable transcripts. Results from RNA FISH (Figure 3.2) reveal RNA distributions that display a poor fit to Poisson-like production of RNA and are more suited to "bursty" switch-like promoters. Another technique labeling stem-loops inserted into the RNA sequence allows for real-time live imaging of RNA synthesis, export and degradation (Figure 3.3). By tracking the number of RNA molecules at a promoter over time, Lenstra *et al.* 2015 found that tracking *GAL10* mRNA with MS2 loops inserted into the sequence exhibited short pulses of transcriptional activity, where multiple polymerases would be engaged at the gene, followed by periods of transcriptional quiescence. This form of gene expression is often termed "bursting", with genes separated into "bursty" or constitutive expression, where the latter can be thought of as a gene exhibiting no quiescent state and constant initiation events. While this technique offers a powerful tool to examine the dynamics of the promoter, particularly in assessing the distributions of nascent RNA molecules at a single locus, recent studies have shown that the introduction of these loops into the sequence can alter the processing and degradation dynamics of the RNA molecules (Garcia & Parker 2015, Garcia & Parker 2016) and even their sub-cellular localisation (Heinrich

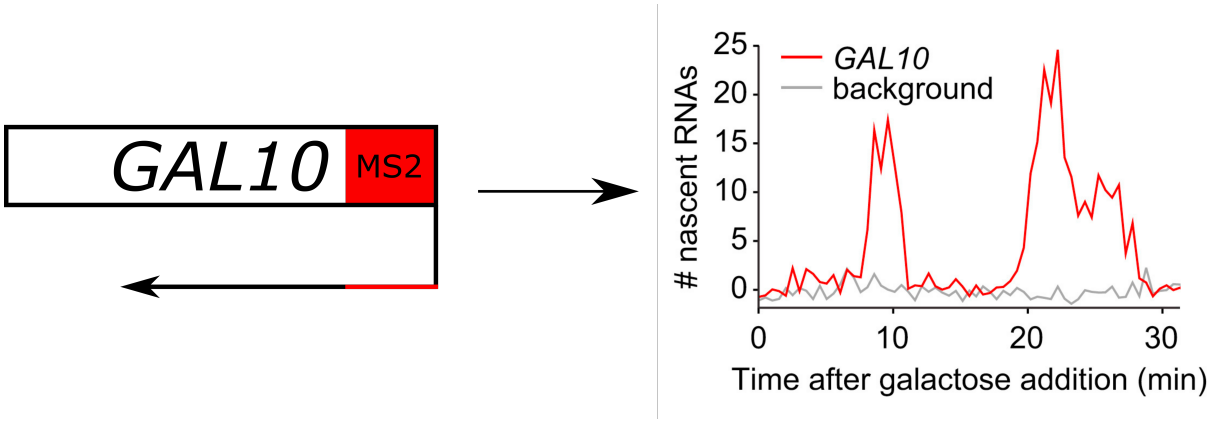


Figure 3.3: Adding MS2 stem loops to the *GAL10* sequence gives a footprint for a tagged MS2 coat protein to bind. Right, a sample trace of *GAL10* initiation, taken as *GAL10* signal in the nucleus. Intensity of signal is proportional to number of polymerases transcribing at one time. Figure adapted from Lenstra *et al.* 2015

et al. 2017). As we wished to combine the results from static RNA distributions and the dynamic results from analysing RNA degradation, we require a system where the two can be directly compared and resemble the endogenous *in vivo* system as closely as possible. Therefore, we chose to continue with the fixed images available from RNA FISH experiments to analyse the full RNA production pathway from promoter to turnover.

A number of possibilities of the exact nature of how these bursts are governed have been proposed. Non-coding RNA has been associated with the control of "leaking" of transcription events (Lenstra *et al.* 2015) as well as the degree of heterogeneity of expression across a population (Huber *et al.* 2016, Steensels & Verstrepen 2016). Alternatively transcription factors such as the SAGA complex have been found to associate with genes more associated with bursty dynamics or population heterogeneity (Zenklusen *et al.* 2008, Ravarani *et al.* 2016).

3.2.3 Polymerase elongation

Polymerase elongation is known to be a complex process in eukaryotic organisms, with the polymerase complex undergoing a number of state transitions varying in dynamics (Duchi *et al.* 2018). Furthermore the potential blockages found in the promoter-proximal region (Ehrensberger *et al.* 2013), around nucleosomes (Churchman & Weissman 2011) and at particular sequences (Larson *et al.* 2014) suggests elongation is not a smooth process.

A more thorough investigation of the possible ranges of polymerase dynamics that exist

will be presented in Chapter 4. Here, however we do not have the resolution of polymerase at different sections of the gene when using RNA FISH. As the signal obtained using ssDNA probes allows for localization of the transcripts to the site of nascent transcription, we are restricted to a more binary view of polymerase engagement being either on or off. Therefore we turned to previously proposed models of polymerase elongation which incorporate stochastic jumps in polymerase from one nucleotide to the next (Choubey *et al.* 2015). Polymerases are modelled as having exponential residence times at each nucleotide, equivalent to a constant rate of elongation in fixed time Δt (Gillespie 1977). The resulting elongation time is thus given by the sum of N exponential distributions, where N is the length of the gene, this follows a Gamma distribution with parameters N and k, corresponding to the elongation rate.

3.2.4 Degradation of RNA

For the majority of genes in *S. cerevisiae* ($\sim$94%), the degradation of the corresponding RNA follows a constant degradation rate, as determined by metabolic labelling of transcripts (Rabani *et al.* 2011, Munchel *et al.* 2011). As such, we model RNA lifetime as following an exponential, where the degradation machinery is assumed to be unsaturated, meaning degradation rates of RNA are linearly proportional to the amount of RNA in a cell.

3.2.5 A model of RNA production

To replicate the dynamics of RNA production and processing we constructed a stochastic model of RNA synthesis, combining aspects from other published models (Peccoud & Ycart 1995, Raj *et al.* 2006, Zenklusen *et al.* 2008, Choubey *et al.* 2015, Choubey 2018, Kumar *et al.* 2015). Building on the "bursty" protein production proposed by Peccoud & Ycart 1995, the production of RNA has similarly been shown not to follow Poisson-like kinetics of the form:

$$\emptyset \rightarrow RNA \rightarrow \emptyset \tag{3.1}$$

in a number of studies (Ietswaart *et al.* 2017, Lenstra *et al.* 2015). Instead, we modeled the promoter as existing in both an "active" and "inactive" state, whereby polymerase

initiates transcription at a constant rate when the promoter is in an active state:

$$Inactive\ Promoter \underset{\beta}{\overset{\alpha}{\rightleftharpoons}} Active\ Promoter \tag{3.2}$$

$$Active\ Promoter \xrightarrow{\gamma} Active\ Promoter\ +\ Initiated\ polymerase \tag{3.3}$$

We then modelled the polymerase as undergoing N steps along the gene, where N is the length of the gene being transcribed. As a polymerase is modelled as moving from one nucleotide to the next with constant rate k, the time until the polymerase moves is distributed as an exponential with rate k. As such, the total time for a polymerase to transcribe a gene, T_{tot} is given by the sum of N exponential distributions with rate k:

$$T_i \sim Exp(k),\ \text{for nucleotide } i \tag{3.4}$$

$$\implies T_{tot} = \sum_{i=1}^{N} T_i \tag{3.5}$$

$$\implies T_{tot} \sim \sum_{i=1}^{N} Exp(k) \tag{3.6}$$

$$\implies T_{tot} \sim Gamma(N, k) \tag{3.7}$$

Once a polymerase has transcribed the full length of the gene, a mature RNA transcript will then be produced instantly, although this can be thought of as a simplification of the true process, which is known to undergo termination, poly-adenylation and packaging into RNPs for export (Mischo & Proudfoot 2013, Aguilera 2005). The RNA transcripts are modelled as undergoing constant degradation, with each RNA having a constant half life. Thus the full system of reactions is (see Figure 3.4):

$$Inactive\ Promoter \underset{\beta}{\overset{\alpha}{\rightleftharpoons}} Active\ Promoter \tag{3.8}$$

$$Active\ Promoter \xrightarrow{\gamma} Active\ Promoter\ +\ Initiated\ polymerase \tag{3.9}$$

$$Initiated\ polymerase \xrightarrow{\Gamma(N,k)} RNA \tag{3.10}$$

$$RNA \xrightarrow{\delta} \emptyset \tag{3.11}$$

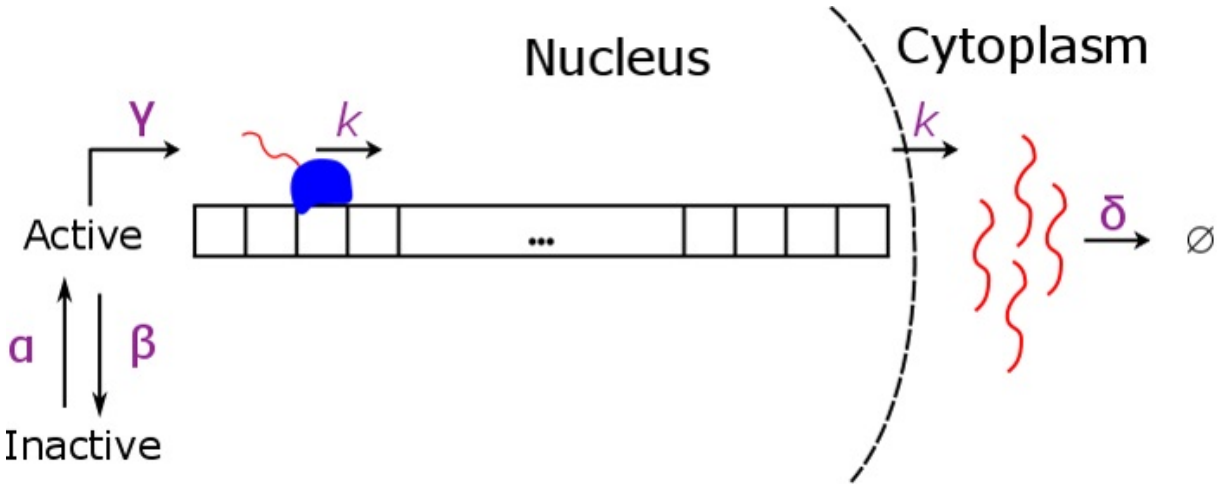


Figure 3.4: A schematic of the model of RNA production. A promoter switches from an inactive to an active state with rate α and vice-versa with rate β. Initiation events occur on an active promoter with rate γ and elongation proceeds at rate k across N nucleotides. A mature, cytoplasmic transcript is produced once the polymerase has elongated across the full gene and is degraded at rate δ.

The system has 5 unknown parameters, activation rate α, inactivation rate β, initiation rate γ, elongation rate k and RNA degradation rate δ and one input, gene length N. Closed-form solutions for the above system have not been found, indeed as the system is not mono-molecular, a simple ansatz does not exist (Jahnke & Huisinga 2007). A solution has been previously presented for a similar, yet simplified system of bursting gene expression kinetics by Peccoud & Ycart 1995 and Raj *et al.* 2006 given by:

$$Inactive\ Promoter \underset{\beta}{\overset{\alpha}{\rightleftharpoons}} Active\ Promoter \tag{3.12}$$

$$Active\ Promoter \xrightarrow{\gamma} Active\ Promoter\ +\ Protein \tag{3.13}$$

$$Protein \xrightarrow{\delta} \emptyset \tag{3.14}$$

The generating function for the distribution of protein in a cell is given by:

$$g(z) = \mathbb{E}\left[z^X\right] \tag{3.15}$$

$$g(z) = {}_1F_1\left(\frac{\alpha}{\delta}, \frac{\alpha+\beta}{\delta}, \frac{\gamma}{\delta}(z-1)\right) \tag{3.16}$$

where X is the probability distribution of protein in a cell and ${}_1F_1$ is the confluent hypergeometric function of the first kind[1]. The mean of the distribution above is given

[1] ${}_1F_1(a,b,c) = 1 + \sum_{k=1}^{\infty} \frac{a(a+1)\ldots(a+k)}{b(b+1)\ldots(b+k)} \frac{c^k}{k!}$

by:

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}\frac{\gamma}{\delta} \tag{3.17}$$

where the fraction $\frac{\alpha}{\alpha+\beta}$ is the probability of a promoter being in an "active" state and the fraction $\frac{\gamma}{\delta}$ is the production-degradation ratio. For a full derivation of the above result see Section 7.2. The results regarding the moments of the system show that one cannot derive the values of every parameter, but are all given relative to each other, e.g. relative to the degradation rate, δ. Furthermore, the dynamics of the promoter are given by the observed initiation rate. As transcription can only initiate when the promoter is in an "active" configuration, the observed initiation rate becomes the quantity:

$$\frac{\alpha\gamma}{\alpha+\beta} \tag{3.18}$$

In the limit $\beta = 0$, this describes the non-bursting Poisson-like dynamics, where the observed initiation rate will be equal to the true rate of polymerase binding the promoter.

3.2.6 Gillespie simulation of RNA distributions

We simulated the distributions of nuclear and cytoplasmic RNA by incorporating the Gillespie stochastic simulation algorithm (SSA) (Gillespie 1977) and the delayed stochastic simulation algorithm (DSSA) (Barrio *et al.* 2006) with the framework described above. Each cell in the population was simulated for 500 minutes worth of reaction time, then the number of RNA molecules in the nucleus and cytoplasm recorded for 10,000 cells at the end of the simulation time. Each reaction in the system (3.8) occurs with constant probability and has exponential waiting time, apart from the elongation reaction. As such, after each initiation event, we gave each polymerase a time to finish elongating, with waiting time given by a Gamma distribution (the sum of N independent exponential random variables with parameter k). We used these finish times in the DSSA framework of Barrio *et al.* 2006, allowing other reactions to occur until the system reaches the next elongation finishing time.

Pseudo-code describing the algorithm used is given in Section 7.3 and CUDA scripts are available at `https://github.com/tbrown91/cuda_paramSearches`.

3.2.7 Parameter estimation via Approximate Bayesian Computation

To obtain estimates for the true parameter values producing the RNA distributions, 1,000,000 sample distributions were simulated for separate sets of parameters. The sets of parameters were sampled via a latin-hypercube (Mckay *et al.* 1979). This sampling method allows for an approximate coverage of the whole sample space evenly, analogous to a uniform prior distribution over all possible values. For each parameter set, the Kolmogorov-Smirnov statistic, defined as the maximum distance between the two cumulative distribution functions, was calculated for the fit to the raw nuclear and cytoplasmic RNA. The best 1,000 (0.1%) parameter sets as defined by their Kolmogorov-Smirnov statistics were then taken as the best-fit parameter sets for the nuclear and cytoplasmic distributions of RNA. The cytoplasmic RNA distribution is independent of the elongation rate k, elongation being analogous to intermediary queues between initiation and RNA degradation. As such the cytoplasmic parameter set define the ratio:

$$\frac{\alpha\gamma}{(\alpha+\beta)\delta} \tag{3.19}$$

or the mean production rate relative to the degradation rate. Similarly, the nuclear RNA distribution is independent of the cytoplasmic degradation rate and defines the ratio:

$$\frac{\alpha\gamma}{(\alpha+\beta)k} \tag{3.20}$$

or mean production rate relative to the elongation rate. Once the best 1,000 parameter sets for the nuclear and cytoplasmic distributions have been obtained, we find the distributions of the posterior relating to the ratios in Equations 3.19 and 3.20. The modal point of the posterior distribution was used to define the rate for each strain (See Figure 3.5 for workflow).

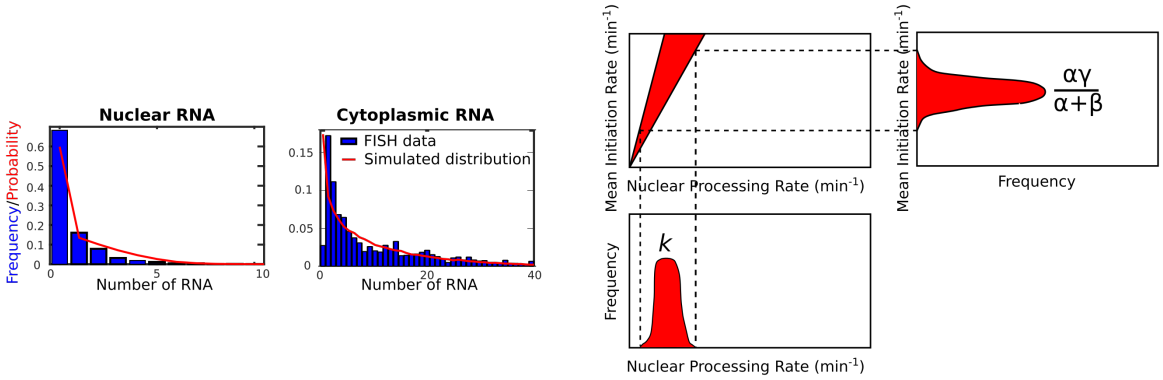


Figure 3.5: A schematic outline of the inference of gene expression parameters from FISH data and modelling. Shown (left) are example distributions of Nuclear and Cytoplasmic RNA from an RNA FISH experiment and the corresponding best-fit distributions from simulations. The best-fit parameters for the nuclear distribution define the ratio of mean initiation rate to nuclear processing rate (right), while the best-fit cytoplasmic parameters give the value of the mean initiation rate. Using the inferred mean production rate from the cytoplasmic distribution combined with the ratio of initiation to processing from the nuclear distribution, we are able to infer the value of the nuclear processing rate.

3.2.8 Determining degradation rates

As highlighted above, modelling of the FISH experimental data only allows us to determine the ratios of mean production rate to degradation rate (Equation 3.19) and mean production rate to nuclear processing rate (Equation 3.20). By finding the degradation rate of the RNA in each strain, we can in turn find the individual values of the mean production rate and elongation rate. We used Northern blotting against the transcript of interest in a transcriptional shutdown experiment for analysis of the *GAL1* mRNA. Cells in galactose-containing media will express the *GAL1* mRNA, however upon transition to glucose-containing media, the *GAL* genes are rapidly shut-off, resulting in exponential decay of the transcripts in the cytoplasm. By measuring the population-wide levels of *GAL1* mRNA at a number of timepoints after switch of media, an exponential curve can be fit to the data to find the approximate degradation rate. Fitting was performed via linear regression to transformed datapoints to find the decay rate δ:

$$R(t) = R_0 exp(-\delta t) \tag{3.21}$$

$$\implies log(R(t)) = log(R_0) - \delta t \tag{3.22}$$

We then used Root Mean Square Error (RMSE) of the linear regression fit to determine the error associated with the inferred parameter δ.

3.3 Transcription of *GAL1-10* locus in glucose and galactose media

The *GAL1-10* locus in *S. cerevisiae* encodes for two key proteins involved in the metabolism of galactose as a carbon source. Under normal growth conditions in media containing glucose, *GAL1* and *GAL10* are not expressed and both transcription units have active antisense transcription. The *GAL1* antisense transcription unit initiates in the terminator of *GAL1* and reads into the bi-directional promoter of *GAL1* and *GAL10*. The *GAL10* antisense transcription unit originates in the body of *GAL10* and is regulated by the transcription factor Reb1. The *GAL10* antisense transcription unit has two isoforms, a short form reading into the bi-directional promoter and a long form (*GAL10*-lncRNA) that is transcribed in a sense orientation to *GAL1*, through the respective gene body (see Figure 3.6).

Upon shift of media to galactose, the corresponding non-coding antisense transcription units are turned off and the sense genes are activated by Gal4. While the role of the *GAL10*-lncRNA on the regulation of the *GAL1-10* locus has been investigated (Lenstra *et al.* 2015), less is known about the role of the *GAL1* antisense transcription unit on the surrounding sense genes. This locus is of particular interest, presenting an example of a typical symmetric sense-antisense transcription layout seen at a large number of genes across yeast and human genomes (Murray *et al.* 2015, Brown *et al.* 2018). To analyse the effects of *GAL1* antisense transcription on the sense expression, we engineered two strains, one with increased levels of *GAL1* antisense transcription and one with much lower levels. Scrambling the TATA box in the engineered terminator resulted in much lower levels of antisense transcription across *GAL1* (Figure 3.7), yet the proportion of A/T to G/C nucleotides remains the same (Murray *et al.* 2015).

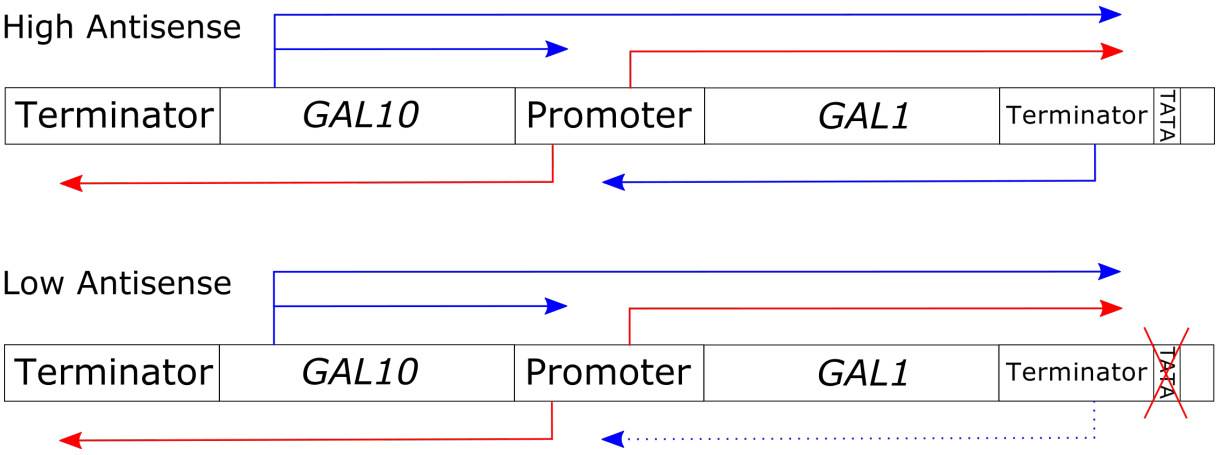


Figure 3.6: Architecture of the *GAL1-10* locus in engineered strains with high and low levels of antisense transcription produced from the terminator of *GAL1*. Sense or protein-coding transcription units are coloured red, while non-coding antisense transcription units are in blue.

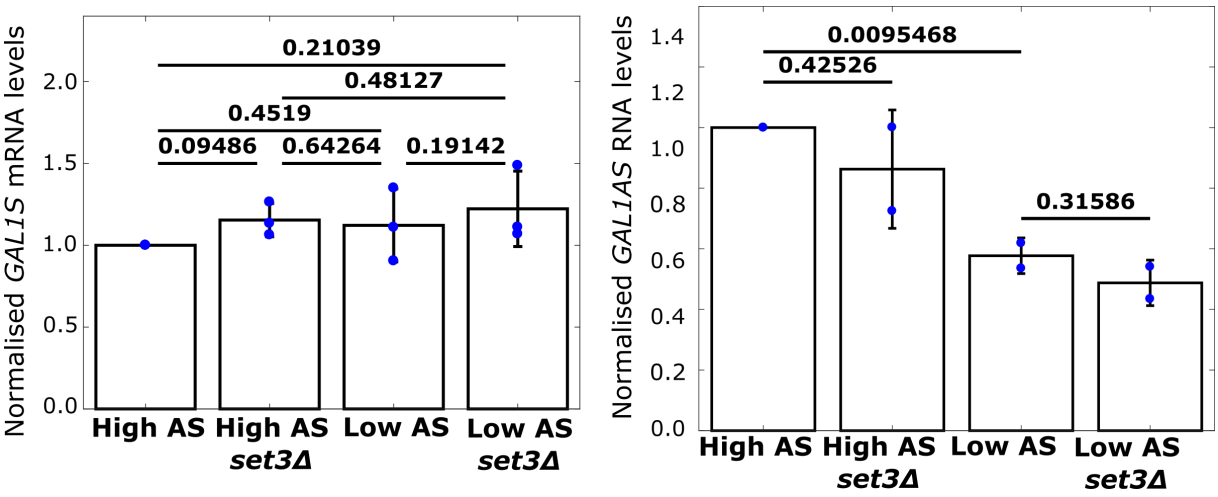


Figure 3.7: Levels of *GAL1* sense and antisense RNA as quantified by Northern blot for the strains in Figure 3.6. Experiments performed by Françoise S. Howe. Taken from Brown *et al.* 2018.

3.4 Does antisense transcription affect sense transcription dynamics?

In order to first investigate the effect of altering levels of antisense transcription on sense transcription, the levels of *GAL1* sense mRNA were calculated via northern blotting (Figure 3.7). Remarkably, although the low AS strain exhibited lower levels of *GAL1* antisense transcript, there appeared to be no effect on the levels of the *GAL1* sense transcript and certainly no increase in levels upon loss of the antisense transcript, as was hypothesised in Eser *et al.* 2016. This result may not be too suprising, as it has been speculated previously that antisense transcription has a role in regulating the levels of

"gene expression noise" (Steensels & Verstrepen 2016), defined as the level of protein variance relative to the mean, rather than the absolute levels of protein (Sevier *et al.* 2016). To further investigate the effect on the distribution of expression of the sense gene, we performed RNA FISH, probing for the sense *GAL1* mRNA transcript.

Coupling population-wide experiments analysing the rate of decay of the sense transcript (Figure 3.11) with stochastic modelling of the RNA distributions (Figures 3.8a, 3.8b) gives the rates of mean production, nuclear processing (Figures 3.10a, 3.10b) and RNA degradation. In the presence of antisense transcription, we found that the corresponding sense transcript was more stable, with a mean half-life of 20 minutes compared with a half-life of ∼14 minutes for the strain with low levels of antisense transcription (Figure 3.11). Furthermore, we found that the mean production rate and nuclear processing rate were both slowed in the presence of antisense transcription (Figure 3.12), potentially confirming a buffering mechanism between initiation and degradation previously observed in *S. cerevisiae* (Sun *et al.* 2013, Haimovich *et al.* 2013).

The Cramer and Choder labs highlighted Xrn1 as a potentially key factor involved in the coupling of transcript production and turnover. We would further suggest here that the chromatin environment set up by processes such as antisense transcription may also be involved in regulating this cross-talk between activity in the nucleus and cytoplasm in RNA surveillance. By having these surveillance mechanisms in the cell, it enforces the idea that all of these processes do not occur in isolation, but exist as part of a gene expression network, with cross-talk from the chromatin, to the transcription machinery to the cytoplasm.

Furthermore, we tested whether the engineered *GAL1* strains used fell into the category of genes which did not follow bursting kinetics. We performed identical model fits to the data with a model where the promoter was constitutively in an active state (equivalent to setting promoter off rate to zero). We observed poor fits to the data (Figure 3.9), suggesting that the long tails of the distributions cannot be accounted for by Poissonian dynamics of promoter activation. This indicates that expression of *GAL1* must be via a promoter with at least 2 states, and cannot be explained by a constitutive promoter that is active with constant rates of initiation. Bi-stable promoters have been shown via

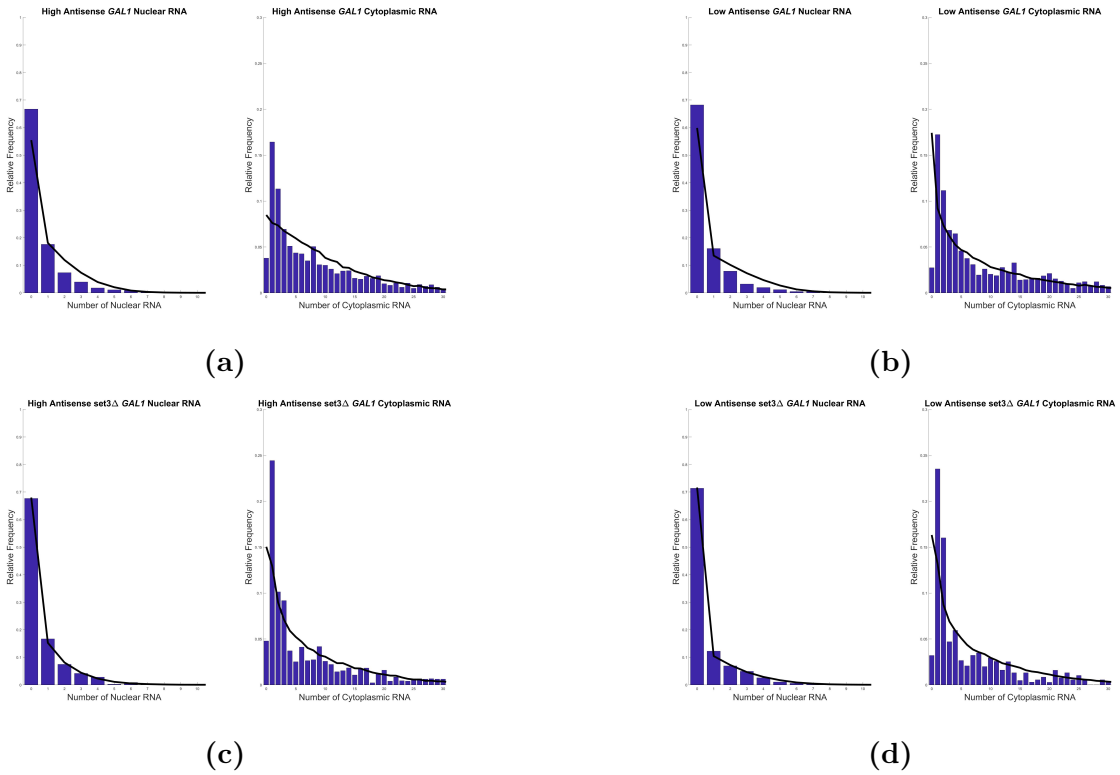


Figure 3.8: Nuclear and Cytoplasmic *GAL1* RNA distributions for engineered *GAL1* locus. Shown are the strains engineered for high (3.8a) and low (3.8b) levels of antisense transcription and upon deletion of histone de-acetylase Set3 (3.8c & 3.8d) Fits from stochastic model are shown as a black line. Experiments performed by Françoise S. Howe. Taken from Brown *et al.* 2018.

live-cell and static imaging at *GAL10* and other genes in *S. cerevisiae*, suggesting this may be ubiquitous (Lenstra *et al.* 2015, Zenklusen *et al.* 2008).

These observations suggest that antisense transcription does indeed affect the dynamics of sense gene expression, without affecting the population-averaged levels of RNA. In order to determine more mechanistic detail of how antisense transcription may affect sense expression, we turned to the effects of antisense on chromatin given its previously recognised relationship with the post-translational modifications of the histones (Murray *et al.* 2015, Marquardt *et al.* 2014).

3.5 Altering acetylation levels mimics effects of altering antisense transcription levels

It has been previously reported that the levels of antisense transcription at a gene has an effect on the chromatin template transcribed by both the sense and antisense polymerase

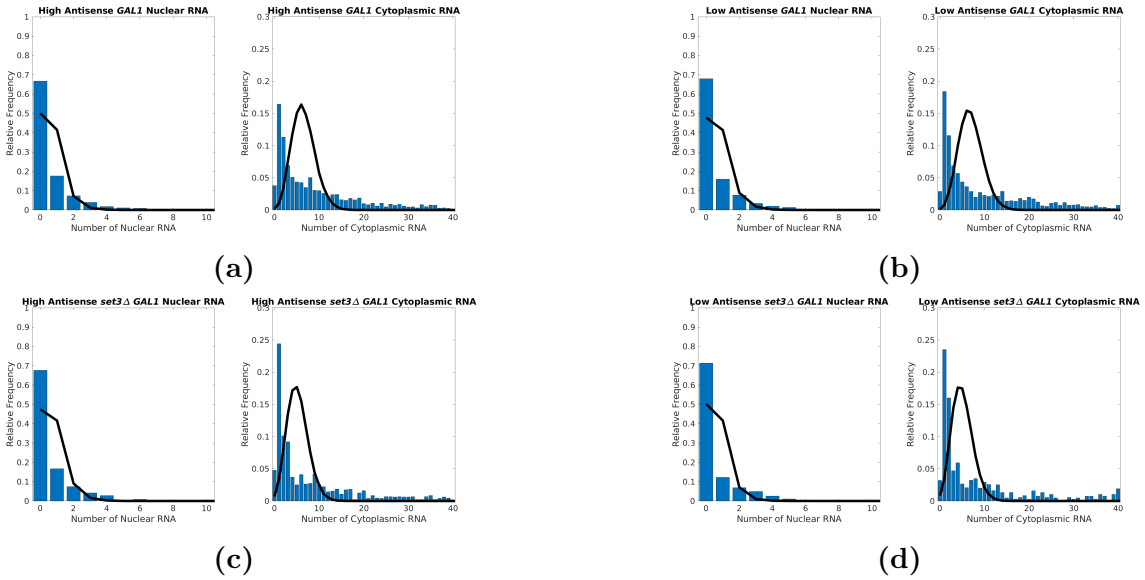


Figure 3.9: Best fits from a model without promoter bursting are shown to the same distributions as in Figure 3.8. Experiments performed by Françoise S. Howe.

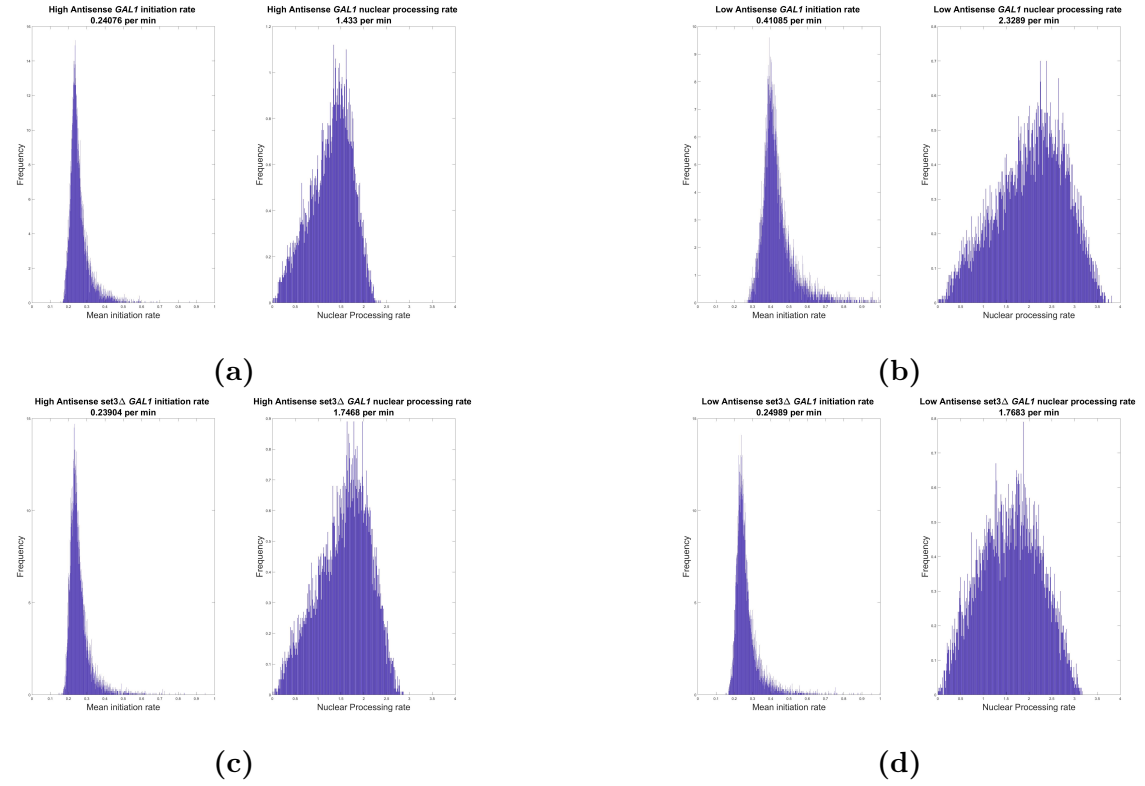


Figure 3.10: Inferred mean production and nuclear processing rates for distributions in Figure 3.8. Shown are the distributions of the best parameter values with the modal value as the title of each graph.

(Murray *et al.* 2015, Brown *et al.* 2018, Marquardt *et al.* 2014). In particular, it has been shown that genes with higher levels of associated antisense transcription exhibit more acetylated histones and lower levels of methylated histones. In order to determine whether the increased acetylation at *GAL1* in the high antisense strain (Murray *et al.*

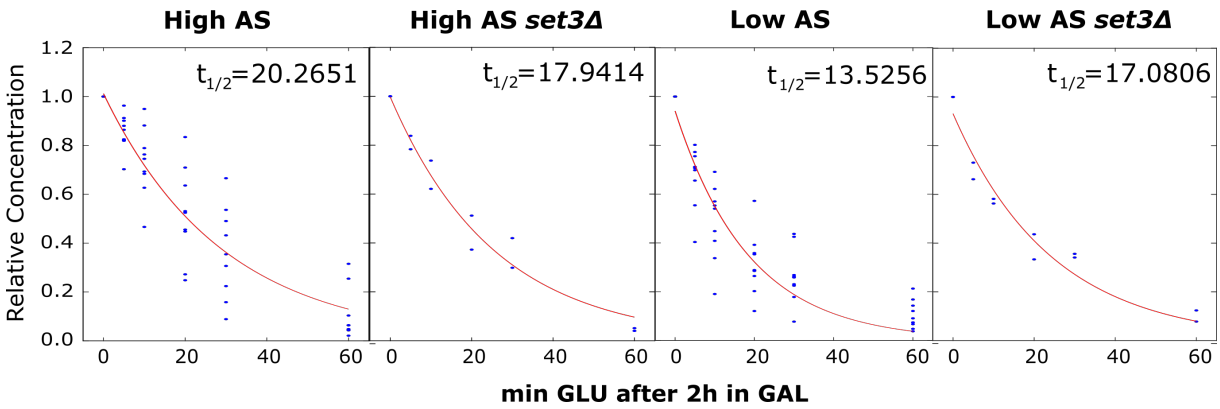


Figure 3.11: Quantified Northern blots showing the levels of *GAL1* sense mRNA upon transcriptional shutdown by shifting cells from glucose to galactose media. In red is shown the exponential fit to the decay curves and inferred half lives. Experiments performed by Françoise S. Howe. Taken from Brown *et al.* 2018.

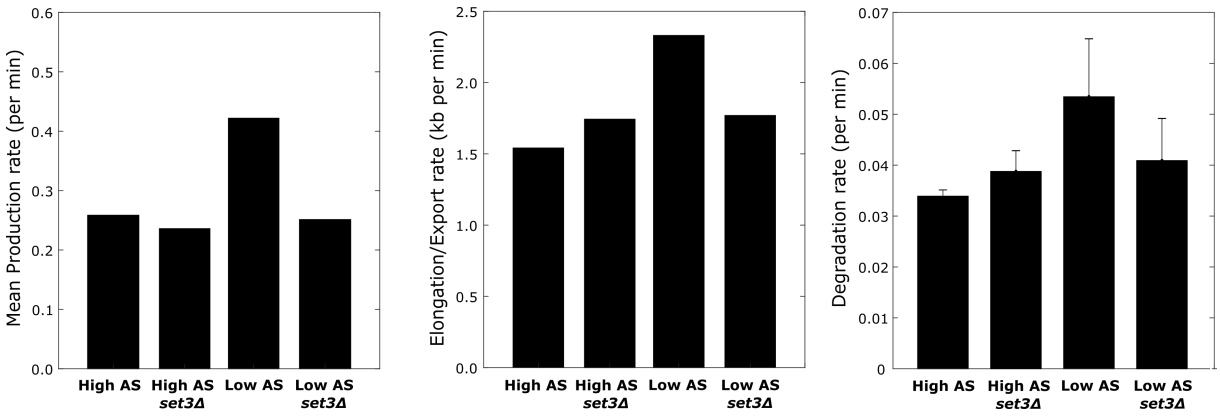


Figure 3.12: Inferred rates from simulations of FISH data in Figure 3.8 and fitting of decay curves in Figure 3.11. Error bars for degradation rates are SEM.

2015) could be the cause of the altered sense expression dynamics, we deleted the histone de-acetylase Set3 from both the low and high antisense strains and repeated the FISH (Figures 3.8c & 3.8d) and *GAL1* shutdown experiments (Figure 3.11). To see if the deletion of Set3 altered the levels of *GAL1* sense and antisense RNA, northern blots for the corresponding RNA were performed after 2 hour's exposure to galactose (sense mRNA), or grown in glucose media (antisense RNA) (Figure 3.7). Set3 is the core component of the Histone De-acetylase complex Set3C, which when deleted results in genome-wide increases in acetylation levels (Weinberger *et al.* 2012, Brown *et al.* 2018).

The levels of both sense and antisense RNA were not affected by the deletion of Set3, suggesting any changes to sense gene expression dynamics would be through the effect of changes to chromatin rather than altering the level of antisense. We found that removing the histone de-acetylase Set3 from the low antisense strain, increasing

the acetylation levels of the *GAL1* chromatin template, mimicked the changes to *GAL1* sense dynamics, decreasing the mean production rate and elongation/export rate while stabilising the transcript in the cytoplasm (Figure 3.12). Furthermore, by deleting Set3 in the high antisense strain, we found that the dynamics of *GAL1* expression were not altered, suggesting a buffering role of antisense transcription to the effect of a loss of Set3. This result has been further confirmed on a genome-wide scale; analysing the change in acetylation levels upon deletion of Set3 found the effect was much more observed at genes without an antisense transcript than those with an antisense transcript (Brown *et al.* 2018).

3.6 Do histone methyltransferases have roles in governing transcription dynamics?

3.6.1 Genes exhibit a range of transcription dynamics

The analysis carried out so far has been on engineered loci, so created to analyse the effects of antisense transcription. We wished to further investigate the range of transcription dynamics across the yeast genome, therefore we turned to the analysis of native genes. 12 genes were separately tagged with GFP at the 3' ends of the genes and FISH probes designed to target the GFP sequence. We thus collected nuclear and cytoplasmic RNA distributions for the 12 genes and the best parameter fit for each gene was again found via the same method as described above (Figures 3.13 - 3.16). Degradation rates for each gene were taken from genome-wide analysis of RNA half-lives in Holstege *et al.* 1998.

The genes chosen displayed a wide array of RNA distributions in the nucleus and cytoplasm. As with *GAL1*, the majority of genes displayed few nuclear transcripts, however the distribution of cytoplasmic RNA ranged from a consistent small number (e.g. *GCV3*, Figure 3.14a) to a widely varying population with cells exhibiting a full range of expression (e.g. *SED1*, Figure 3.16a). The inferred mean initiation and elongation rates show a remarkable correlation, suggesting coupling of transcription promoter activity to the speed of the elongating polymerase (Figure 3.17). The RNA distributions obtained are similar to those seen by Zenklusen *et al.* 2008 for *KAP104* (Figure 3.15c). Therefore

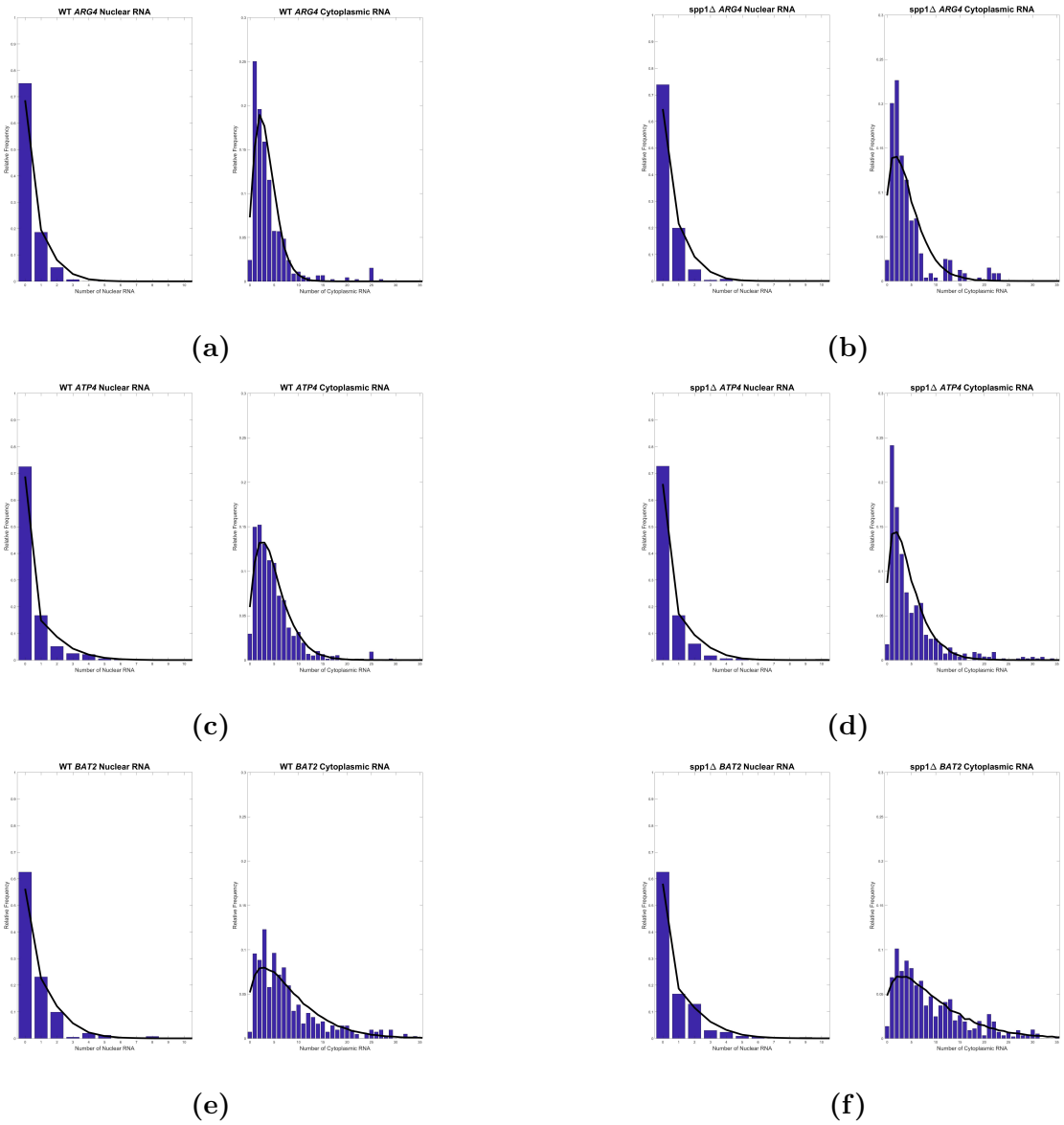


Figure 3.13: Raw nuclear and cytoplasmic RNA distributions from *ARG4* gene in wild-type (3.13a) and *spp1*Δ (3.13b), *ATP4* gene in wild-type (3.13c) and *spp1*Δ (3.13d) and from *BAT2* gene in wild-type (3.13e) and *spp1*Δ (3.13f). Fits from stochastic model are shown as a black line. Experiments performed by Meredith Wouters.

we analysed these parameters in light of the observations in this study.

Unlike the study mentioned, the probes used in this study were located at the 3' end of genes, therefore the signal produced by FISH probes will reflect both the elongation rate through the final ∼700bp of the sequence given by the GFP tag and the export of the RNA from the nucleus. This may also explain discrepancies between nuclear distributions between those published and those presented here. The inferred rates however display a range of rates between the genes studied. Here, we are presented with a system which allows us to examine the effect of Spp1, a regulator of the Set1C histone methyltransferase

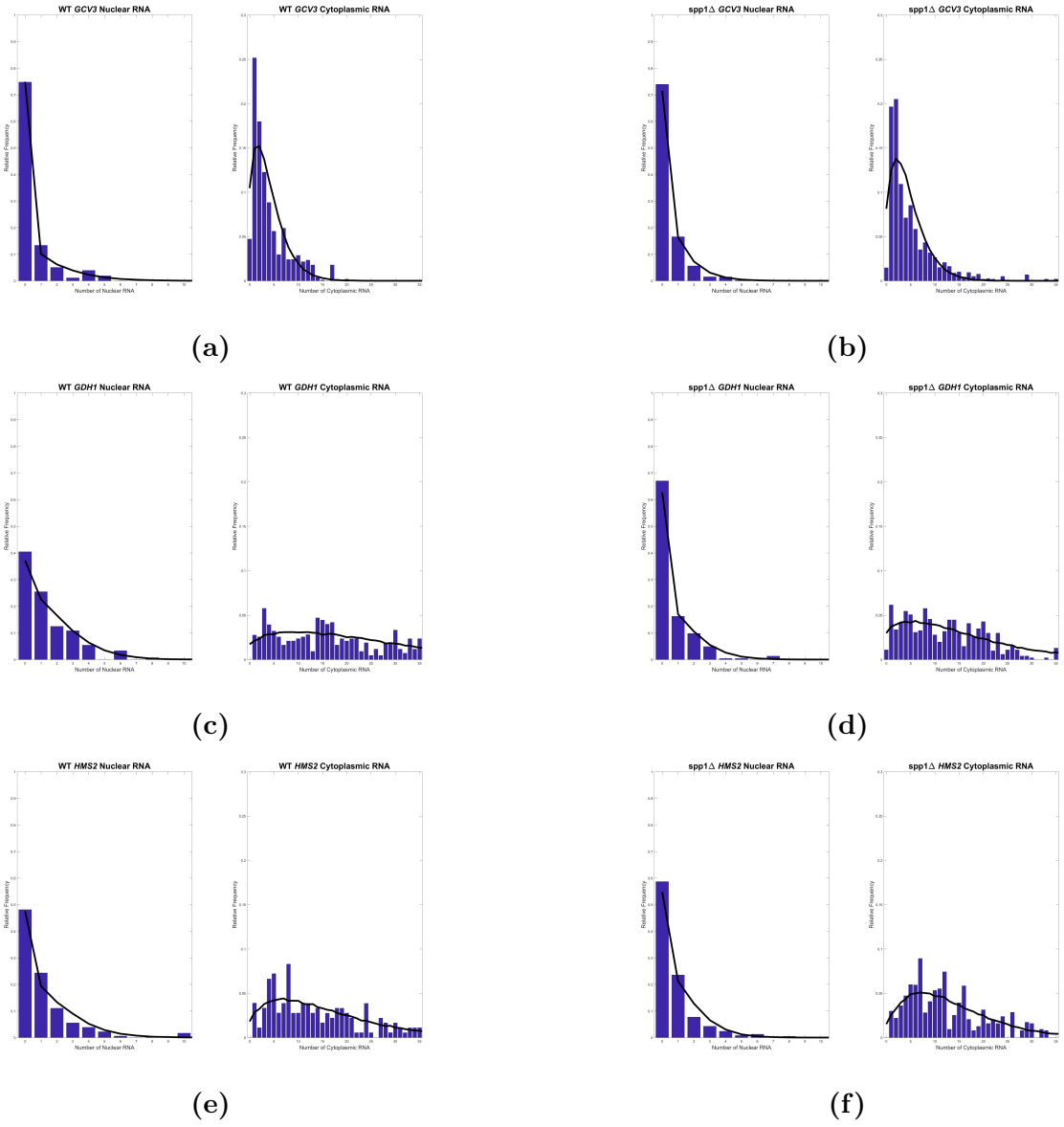


Figure 3.14: Raw nuclear and cytoplasmic RNA distributions from *GCV3* gene in wild-type (3.14a) and *spp1*Δ (3.14b), *GDH1* gene in wild-type (3.14c) and *spp1*Δ (3.14d) and from *HMS2* gene in wild-type (3.14e) and *spp1*Δ (3.14f). Fits from stochastic model are shown as a black line. Experiments performed by Meredith Wouters.

targeting H3K4, on gene expression across the genome.

Taking a reasonably large sample of genes on which to perform RNA FISH, we observe a large range of dynamics from low to high initiation and elongation rates and varying levels of steady-state expression. By performing RNA FISH on wild-type cells and following deletion of *SPP1*, we have the opportunity to investigate the effects of H3K4me3 on gene expression dynamics at genes which exhibit these different dynamics. While we may expect a direct role on gene expression after interfering with H3K4me3 given its annotated role in active transcription (Santos-Rosa *et al.* 2002, Howe *et al.* 2017), here

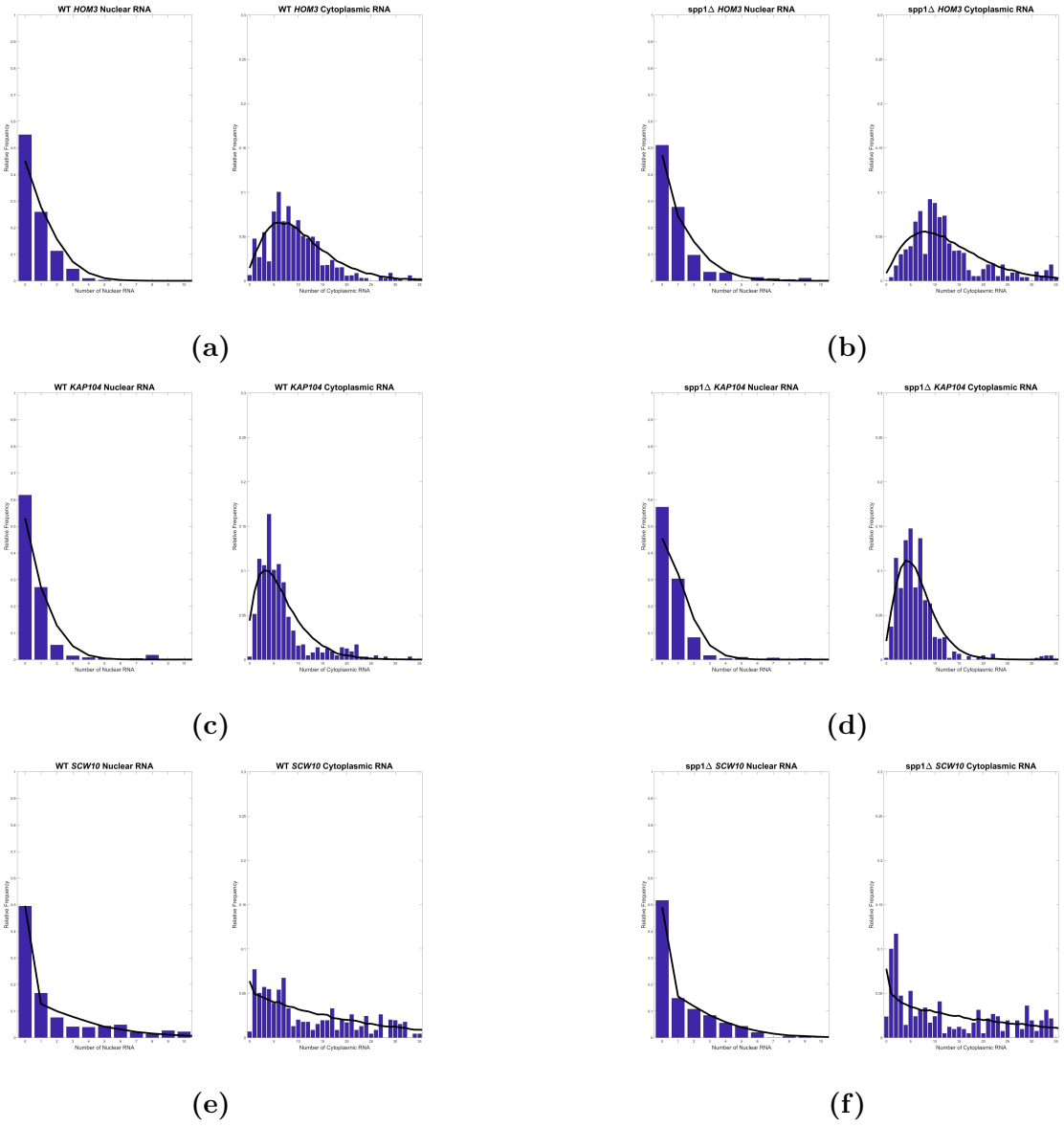


Figure 3.15: Raw nuclear and cytoplasmic RNA distributions from *HOM3* gene in wild-type (3.15a) and *spp1*Δ (3.15b) , *KAP104* gene in wild-type (3.15c) and *spp1*Δ (3.15d) and from *SCW10* gene in wild-type (3.15e) and *spp1*Δ (3.15f). Fits from stochastic model are shown as a black line. Experiments performed by Meredith Wouters.

we can potentially separate genes based on both their wild-type dynamics and response to loss of H3K4me3.

3.6.2 H3K4me3 does not control transcription dynamics

In order to investigate the role of trimethylation of lysine 4 on histone 3 (H3K4me3), a reported marker of active transcription (Santos-Rosa *et al.* 2002, Howe *et al.* 2017), we repeated the FISH experiments on the chosen genes in strains lacking Spp1. Spp1 forms part of the histone methyltransferase complex Set1C and is responsible for only the tri-

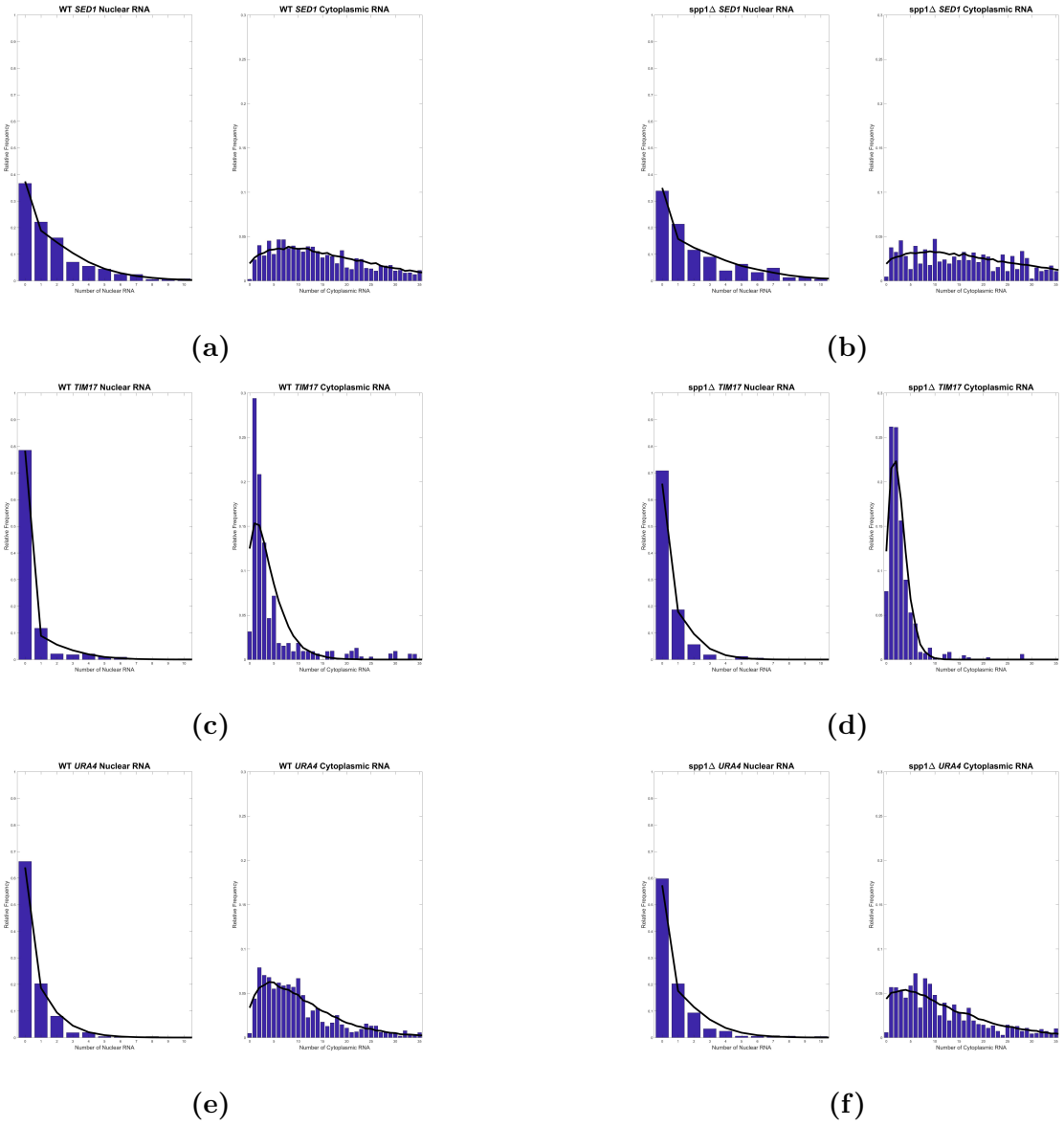


Figure 3.16: Raw nuclear and cytoplasmic RNA distributions from *SED1* gene in wild-type (3.16a) and *spp1*Δ (3.16b), *TIM17* gene in wild-type (3.16c) and *spp1*Δ (3.16d and from *URA4* gene in wild-type (3.16e) and *spp1*Δ (3.16f). Fits from stochastic model are shown as a black line. Experiments performed by Meredith Wouters.

methylation of H3K4 (Figure 3.18). While H3K4me3 levels are highest at genes with most active transcription, the deletion of components of the Set1 complex see little change on expression of genes at the level of nascent or mature RNA (Howe *et al.* 2017). We aimed to further probe the role H3K4me3 may have on the dynamics of gene expression through the framework presented here.

After deletion of *SPP1*, we observed no overall impact on the distributions of nuclear or cytoplasmic RNA at any of the genes analysed (Figures 3.13 - 3.16). This does not immediately indicate anything about the underlying dynamics, but does highlight that

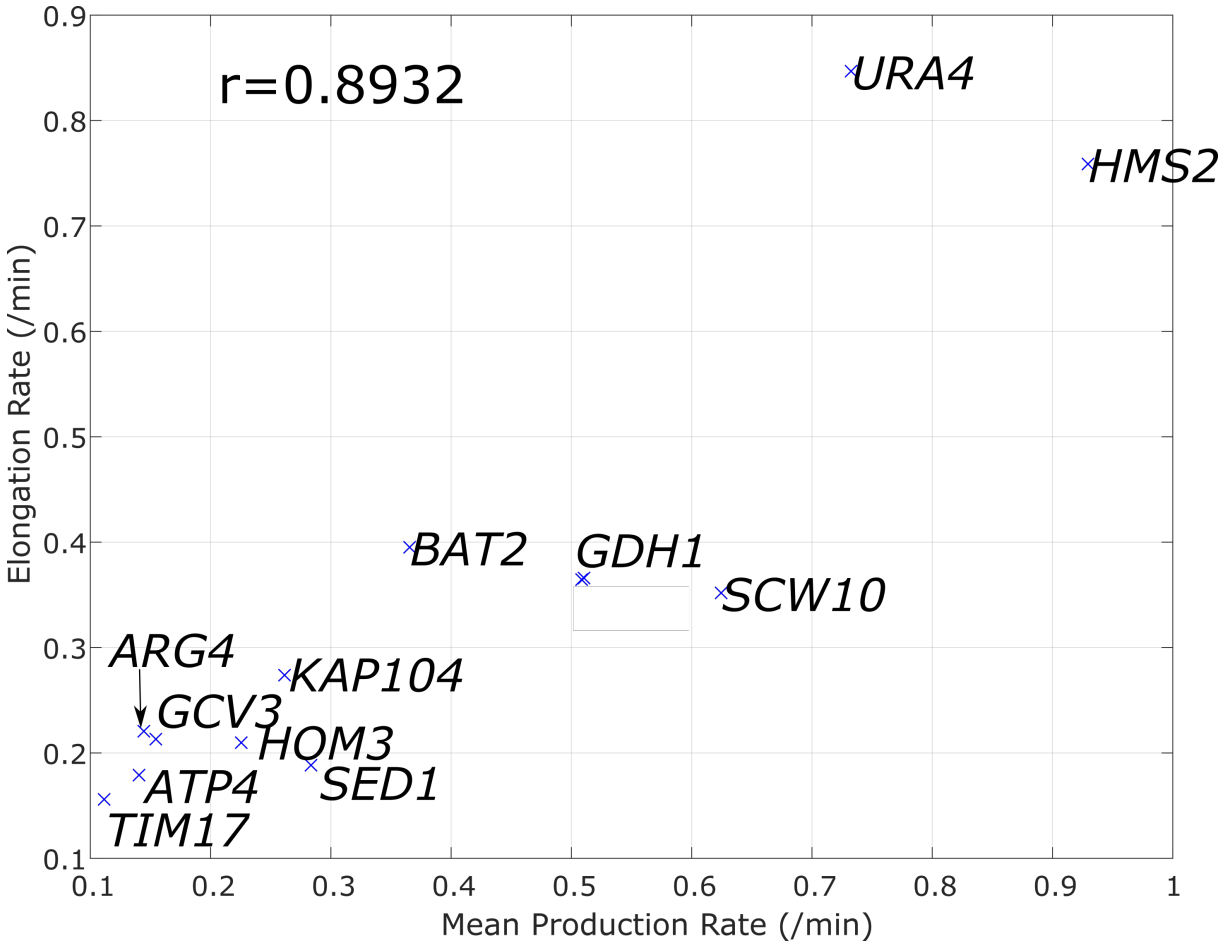


Figure 3.17: Inferred mean initiation and elongation rates for the 12 analysed endogenous genes.

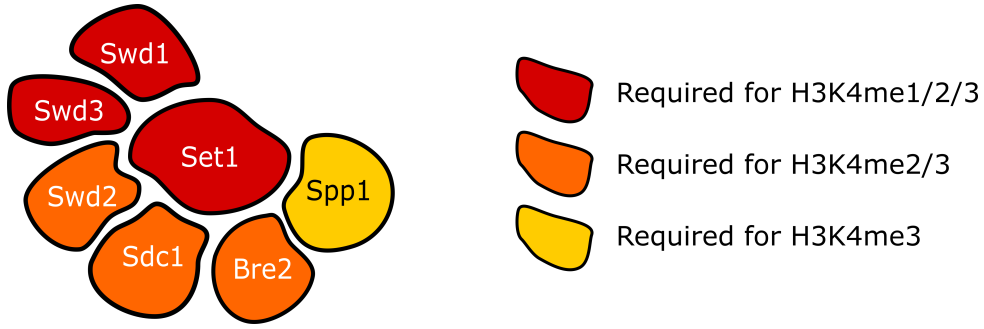


Figure 3.18: Components of *S. cerevisiae* methyltransferase Set1 complex. Highlighted are the roles of each subunit regarding methylation of lysine 4 on histone 3. Adapted from Howe *et al.* 2017.

even though we observed a wide range of RNA distributions in wild-type cells, there are not any obvious changes to the distributions which are generally reflective of ranges in gene expression or noise (Weinberger *et al.* 2012).

Separating the dynamics into initiation and nuclear processing, we observed no large or consistent changes in either rates upon deletion of Spp1 (Figure 3.19). Almost all genes studied exhibited no changes in either their initiation rates, barring a decrease of initiation

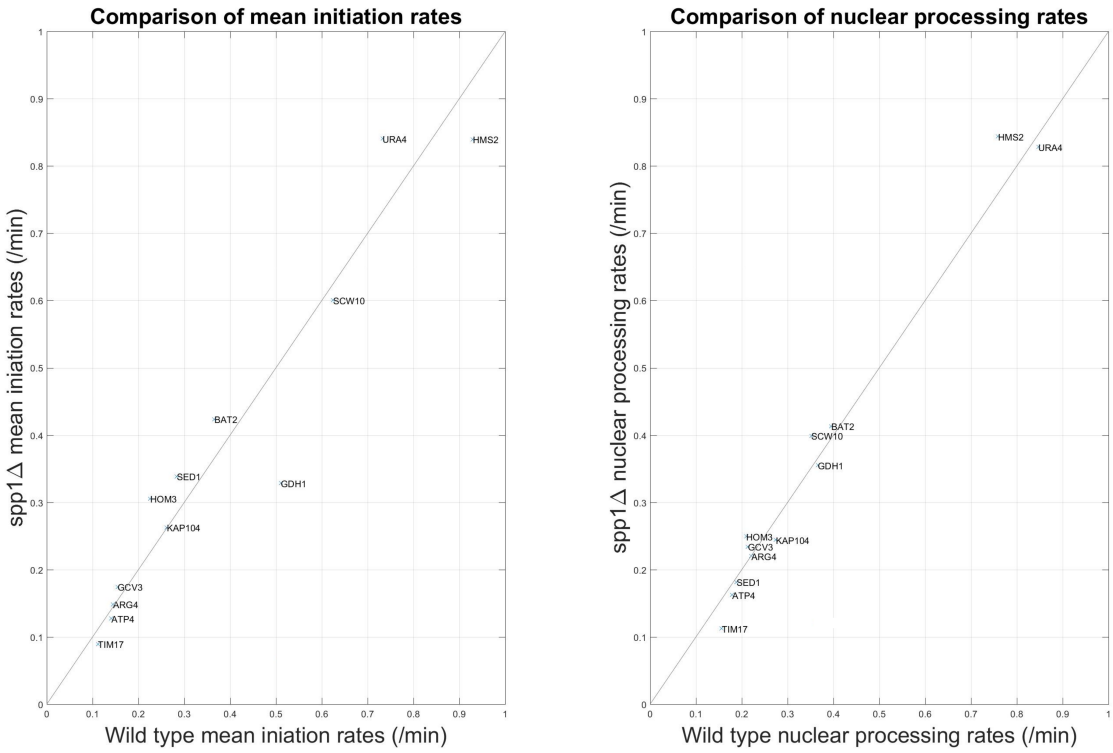


Figure 3.19: Comparison between rates of mean production and nuclear processing rates for 12 genes in wild-type and upon deletion of Spp1.

rate at *GDH1* and no changes at any genes to the nuclear processing rate. Expression of *GDH1* has been shown to be media-specific, controlled by transcription factors which are nitrogen or carbon-metabolism regulated (Riego *et al.* 2002). It may be that the removal of Spp1 from the cell disrupts the metabolism of glucose in *S. cerevisiae*, resulting in altered activity of these factors, whether this arises through disruption of the chromatin or to cell division as part of an alternative complex (Karányi *et al.* 2018, Adam *et al.* 2018).

3.7 Conclusions

3.7.1 Antisense and histone acetylation affect rates of transcription

Here we have shown that single-cell, single-molecule experiments tracking the number and location of RNA molecules from individual genes allow us to determine the promoter, gene body and degradation dynamics over the life-time of RNA molecules from synthesis

to turnover. Beyond studying the dynamics of endogenous genes as in Zenklusen *et al.* 2008, we show that a powerful use of this approach is to investigate the effect of altering the surrounding chromatin environment on the rates of transcription of a gene. Using strains with altered levels of antisense transcription across the *GAL1* locus (Murray *et al.* 2015, Brown *et al.* 2018) gave us the opportunity to investigate the effect antisense transcription has on the expression of the corresponding sense gene.

We found that while steady-state levels of the sense mRNA remain constant in the presence of high or low levels of antisense transcription, the role of antisense transcription is to modulate the rates at which these transcripts are produced and degraded in the cytoplasm. We found that with high levels of antisense transcription, both transcription initiation events and the elongation/export phases of transcript production were slowed (Figure 3.12). However, we found by performing cell-population experiments tracking levels of *GAL1* mRNA upon transcriptional shut-off that these decreases in transcription rates were coupled by an increase in the transcript stability (Figure 3.11), reflecting the ability of the cell to buffer the activity of the promoter with the turnover of transcripts. It is perhaps an important distinction that the rates of turnover reflect the activity of the polymerase on the chromatin, rather than antisense transcription directly affecting the stability of the transcription. However, the same observation was also confirmed on a genome-wide scale, as genes with an antisense transcription unit were shown to have a more stable transcript than those without (Brown *et al.* 2018).

It has been previously observed that genes with high levels of antisense transcription are associated with a unique chromatin architecture (Murray *et al.* 2015, Marquardt *et al.* 2014). We wished to determine whether the observed effects antisense transcription had on sense transcription dynamics could be accounted for through the increased acetylation levels observed in the promoter region and gene body. By removing the histone de-acetylase Set3, we created a chromatin template with higher levels of histone acetylation and saw that in the absence of antisense transcription, this recovered the dynamics of the high antisense strain. Furthermore, we observed that in the presence of antisense transcription, removal of Set3 did not affect the dynamics of sense transcription, suggesting an interaction between antisense transcription and Set3 to modulate transcription

dynamics in a compensatory manner.

This highlights the fact that the cell has multiple mechanisms by which to control the levels of post-translational modifications of histones. By removing the histone deacetylase as well as increasing the levels of antisense transcription, the chromatin appears to reaches a threshold of acetylation and, correspondingly, there is no effect on the dynamics of transcription when comparing the individual "mutants" with the combined (Figure 3.12). With these multiple mechanisms working together to modulate the chromatin, it offers a number of compensatory methods by which the cell can maintain a "wild-type" environment under changing conditions for the cell.

3.7.2 Transcription dynamics are distinct at individual genes

A number of endogenous genes were investigated, which display a wide range of dynamics, in particular showing a correlation between the mean initiation rates of transcription events and the rate of nuclear processing (Figure 3.17). While it has been observed that the cell possesses at least one buffering mechanism linking the activity of the promoter to the turnover of transcripts (Haimovich *et al.* 2013), results here suggest that the procession of the polymerase may also be linked to the initiation activity of the promoter.

By removing the H3K4 methyl-transferase Spp1, we observed both that the levels of the RNA and the distributions across a population are not affected by disruption of the levels of H3K4me3 (Figure 3.19). While we do observe a range of RNA distributions and transcription parameters, it appears that these dynamics are not dictated by the activity of Spp1 or the levels of H3K4me3, even though this is a mark of active transcription (Santos-Rosa *et al.* 2002, Howe *et al.* 2017).

This suggests that while Spp1 is associated with transcription, its role may be much more pronounced in chromosome division during meiosis (Karányi *et al.* 2018, Adam *et al.* 2018). This particularly highlights that although transcription must occur through the chromatin environment, there may not necessarily be feedback between the two. It may be the case that certain marks, such as the methylation of H3K4, are a by-product of transcription, whereas others, such as acetylation through the gene, have a more direct impact on the activity of the polymerase.

Chapter 4

Polymerase activity is distinct across the genome

The majority of the experimental data presented here has been previously published by the Weissman, Murphy & Mellor labs. Unpublished data was produced by Harry Fischl, Ülkü Uzun, Shidong Xi and Michael Tellier.

4.1 Introduction

4.1.1 From single genes to genome-wide analysis

The analysis of single-cell RNA FISH datasets allowed us to determine the dynamics of RNA production and processing for a number of genes. However, this process quickly becomes experimentally too intensive when we wish to probe the dynamics of RNA production genome-wide. In particular, we wish to determine whether the dynamics we have observed via RNA FISH and modelling in Chapter 3 are conserved among other genes. Have we only covered a sub-class of genes regarding their transcriptional dynamics or does each gene have distinct transcriptional dynamics?

As we can no longer rely on RNA FISH as a method to analyse transcriptional dynamics on the scale we wish, we must move to a more genome-wide view of obtaining quantitative transcription data. Along with a new experimental framework, we require more sophisticated down-stream processing of the data. Bioinformatic processing of genome-

wide data also prompts the need for a new model of transcription in order to determine underlying dynamics from the obtained data, which will be presented in Chapter 5.

4.1.2 High resolution views of polymerase activity

Methods investigating transcription have recently accelerated in complexity and resolution. Previously, transcription studies have relied on RNA-level quantification at either single-gene or genome-wide levels, or by tracking the levels and location of polymerase on DNA via ChIP-seq, which originally prompted discussion regarding inferring dynamics of polymerase from studying the steady-state location of polymerase (Ehrensberger *et al.* 2013). Recently, more sophisticated techniques have been developed allowing for quantitation of nascent transcription, as opposed to mature nuclear or cytoplasmic RNA, in particular separating out transcription on each strand of the genome, not possible with techniques such as ChIP-seq. The development of more high-resolution techniques investigating transcription introduces the potential for more sophisticated data analysis to unpick the intricacies of the transcription process.

Before we can begin to construct models of RNA PolII elongation, we require a sophisticated method of analysing nascent transcription data and the location of transcribing polymerase as given by Native Elongating Transcript sequencing (NET-seq or mNET-seq, Churchman & Weissman 2011, Nojima *et al.* 2015). Under the NET-seq protocol, Rpb3, a subunit of the protein-coding RNA Polymerase II complex, is FLAG-tagged. The tagged protein is then isolated, and the protected fragments of RNA found in the active site of the polymerase complex are sequenced. The short fragments of protected RNA ($\sim$18bp) allow for a high-resolution map of the location of transcriptionally-engaged polymerase, but taking the most 3' nucleotide of the RNA fragment shows the location of the last nucleotide incorporated into the polymerase (see Figure 4.1). This allows us to determine with single-nucleotide resolution the location of all transcriptionally-engaged RNA PolII complexes in the genome. The mNET-seq protocol aims to replicate this procedure, however does so first through isolation of the chromatin. The polymerase is then released from the chromatin fraction by digestion with MNase, then total, unphosphorylated or CTD phospho-isoforms of the RNA PolII are captured via antibodies (Nojima *et al.* 2015).

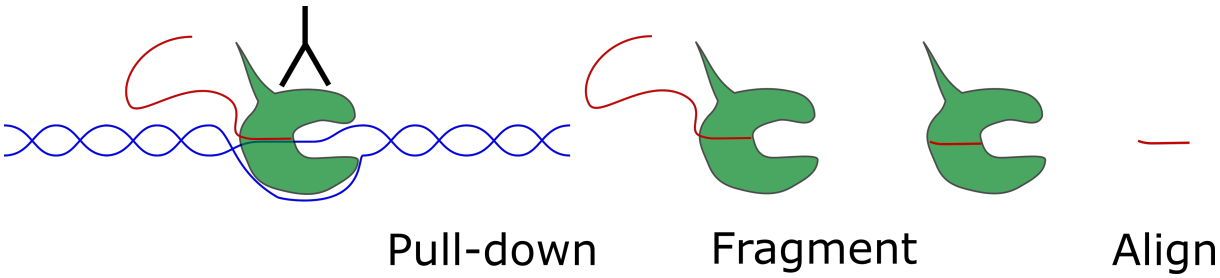


Figure 4.1: An outline of the NET-seq protocol. Tagged RNA Pol II is pulled down, exposed RNA is fragmented, leaving the protected RNA found in the active site of polymerase. The short fragments of RNA are then sequenced and aligned to the genome, revealing the location of the last incorporated nucleotide into the transcriptionally-engaged polymerase.

Currently poorly characterised is the degree of heterogeneity that exists in polymerase profiles across the transcribed genes of *S. cerevisiae* and human cell lines. By developing more sophisticated bioinformatic techniques to analyse nascent transcription, we are able to distinguish sub-classes of genes which exhibit distinct distributions of polymerase across their genes. A further outstanding question is whether these sub-classes of polymerase distributions are functional. It has been long known that some stress-response genes have poised polymerases at the 5' ends of their genes (Rougvie & Lis 1988), however here we show that the number of genes with the majority of polymerase at the 5' end of genes is much greater than previously thought. Furthermore, we ask whether the profile of polymerase is instructional for the amount of downstream RNA, indicative of the true "expression" of a gene. The work in this Chapter will establish the types of output data we observe from NET-seq and mNET-seq experiments, which is of a highly quantitative nature, making it suitable for mathematical modelling, to then understand the nature of the underlying mechanisms and dynamics governing RNA PolII transcription.

4.1.3 The dynamics of RNA polymerase II

While it may be tempting to think of transcription as a smooth process from initiation to termination, the true picture of polymerase dynamics is much more complex. The speed of transcribing polymerase *in vivo* has been reported to range from as slow as 400bp/min (Booth *et al.* 2018) up to well over 10kb/min, or even 50kb/min (Gressel *et al.* 2017, Maiuri *et al.* 2011). Previous studies have indeed highlighted that there may not be one universal speed of polymerase across different organisms (e.g. Zenklusen *et al.* 2008, Booth *et al.* 2018), or even across the genome within the same organism

(Gressel *et al.* 2017). What is not known is what contributes to these different apparent elongation rates and whether these ranges of dynamics are at all indicative of alternative function. The polymerase complex is known to undergo a number of changes from the stages of initiation to termination with modifications to the polymerase molecule itself on the CTD (Zaborowska *et al.* 2016), as well as proteins associating and dissociating with the polymerase complex as it moves through the gene (Fischl *et al.* 2017). A number of the known dynamic transitions which make up the complex process of transcription are highlighted below.

Polymerase stalling, pausing and backtracking

One possible explanation for the range of observed elongation speeds for the same enzyme is the degree of stochastic pausing or stalling by the transcribing elongation complex, which occurs throughout the gene body (Mayer *et al.* 2017). Indeed, it has been observed *in vitro* using high resolution optical tweezers experiments that polymerase stochastically pauses while transcribing naked DNA in nucleotide-limited conditions (Larson *et al.* 2014). The time for polymerase to be released from these pauses in some instances are on a greater order than the time for polymerase to transcribe a ∼1.5kb stretch of naked DNA. Further evidence for stalling and backtracking *in-vivo* has been presented in relation to the backtrack-release activity of Dst1 (TFIIS) in run-on experiments (Sigurdsson *et al.* 2010) or tracking the location of PolII in *S. cerevisiae* (Churchman & Weissman 2011) and *E. coli* (Larson *et al.* 2014). This goes some way to explaining the ease with which discrepancies may arise between the observed elongation rates *in vivo*, as the polymerase is known to transcribe through a number of obstacles including nucleosomes and DNA-binding proteins, which may contribute to additional stalling and backtracking events.

In both *S. cerevisiae* and human cells it is often reported that the polymerase undergoes stalling, pausing and backtracking as part of the elongation process, however it is not known how much of this occurs or whether it is the same at every gene or in different conditions (Larson *et al.* 2014, Chen *et al.* 2015, Churchman & Weissman 2011). We wish to determine here what proportion of time the polymerase spends in a moving state on its journey from the 5' to the 3' end of the gene, how much this differs between genes and

within genes and what controls these dynamics states of the polymerase.

Promoter-proximal pausing

Polymerase stalling or pausing is proposed to have functional roles beyond stochastic noise to polymerase passage. The slowing or stopping of polymerase after initial rounds of initiation has been speculated as a place for the elongation complex to undergo a checkpoint, determining, for example, whether the polymerase is associated with the correct elongation factors, or post-translational modifications (Erickson *et al.* 2018). Immediately downstream of the TSS, polymerase has been shown to pause in a primed position (Rougvie & Lis 1988), before proceeding into the gene body and processive elongation. More recently, this step of transcription has been termed the "Early Elongation Checkpoint" (EEC, Tellier *et al.* 2016), and is associated with a similar pause-like checkpoint found in the termination regions of genes (Laitem *et al.* 2015).

In metazoa, negative elongation complexes DRB-Sensitivity Inducing Factor (DSIF) and Negative Elongation Factor (NELF) associate with polymerase after initiation and are thought to bring about the downstream pause. The hSPT5 subunit of DSIF is then phosphorylated by CDK9, part of positive elongation factor P-TEFb (Marshall & Price 1995, Wada *et al.* 1998b, Kim & Sharp 2001), converting DSIF from a negative to a positive elongation factor and releasing NELF from the elongation complex (See Figure 4.2). NELF and the unphosphorylated form of DSIF are reported to slow the elongating speed PolII, while the phosphatase activity of P-TEFb increases the elongation speed after the initial period of slower polymerase elongation (Zhou *et al.* 2012). Recently, run-on analysis of polymerase with inhibited Cdk9 activity in *S. pombe* suggests that, at the 5' end of genes, polymerase moves much slower than in the gene body and this transition is determined by the activity of Cdk9 in phosphorylating Spt5 (Booth *et al.* 2018).

It is still not fully understood whether P-TEFb acts on all genes, especially at steady-state, and the kinetics of release from early into mid-transcription elongation. Investigating the distributions of transcribing polymerase before and after inhibition of CDK9 in mammalian cells we are able to determine the genome-wide effects of the phosphorylation of DSIF and the time-scale over which CDK9 acts.

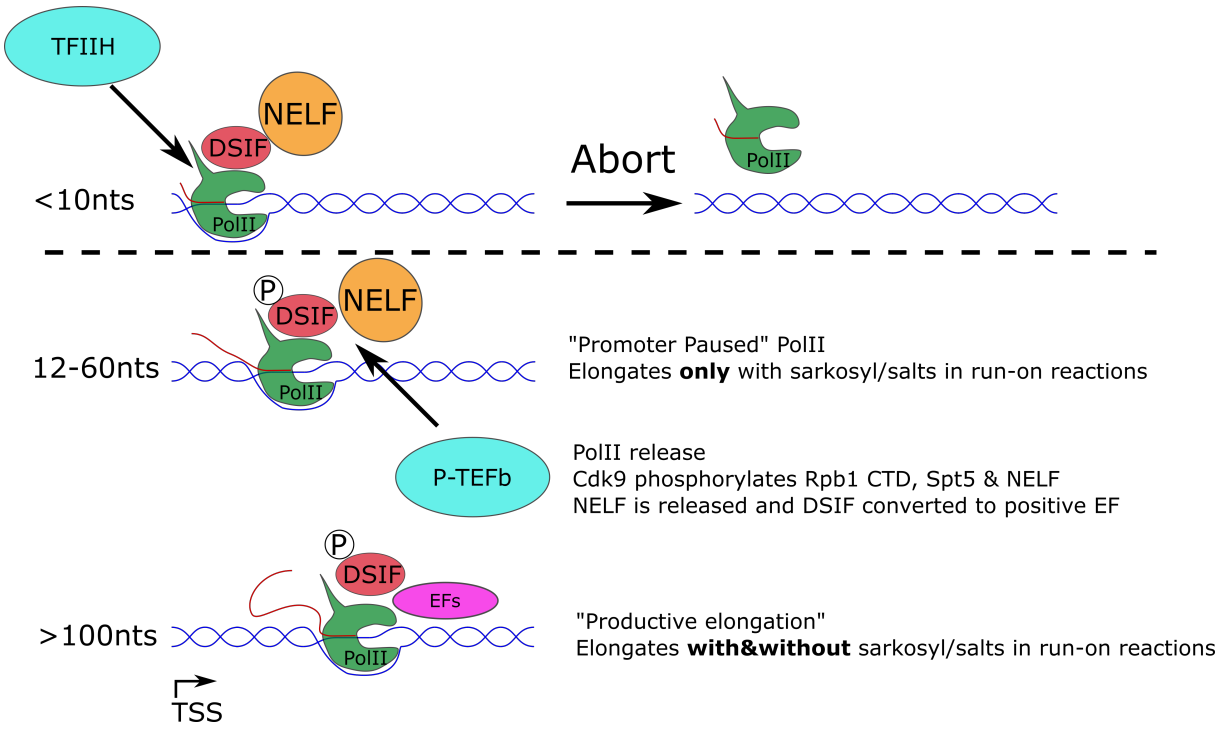


Figure 4.2: A schematic of the steps involved in promoter proximal pausing and release into processive transcription. TFIIH phosphorylates serine 5 residues of the polymerase CTD at initiation, while negative pause factors DSIF and NELF are recruited to the polymerase. p-TEFb phosphorylates Spt5 of the DSIF complex, converting DSIF to a positive elongation factor and removing NELF from the transcription complex. Positive elongation factors are then recruited to the polymerase as it transcribes through the gene body.

4.1.4 Aims

In this Chapter we aim to develop and employ bioinformatic tools to assess the degree of variation in polymerase activity across the *S. cerevisiae* and human genomes. Using NET-seq as a metric of polymerase activity, we aim to determine how the polymerase is distributed across a gene while transcribing. It is not known what functional roles distinct distributions of polymerase may play in the grand scheme of gene regulation, but a better idea of the distributions of polymerase across genomes and organisms will point the way for a more in-depth understanding of transcription dynamics in eukaryotic organisms.

Furthermore, we are able to determine what roles known and speculated elongation regulatory factors and histone writers and erasers have on the distributions of polymerase. By either removing the corresponding protein from the cell or by inhibiting its activity, we are able to infer the role of these factors on transcription elongation by analysing the distributions of transcribing polymerase after inhibition/deletion. While many of the factors presented here have been previously annotated, we present here the degree of

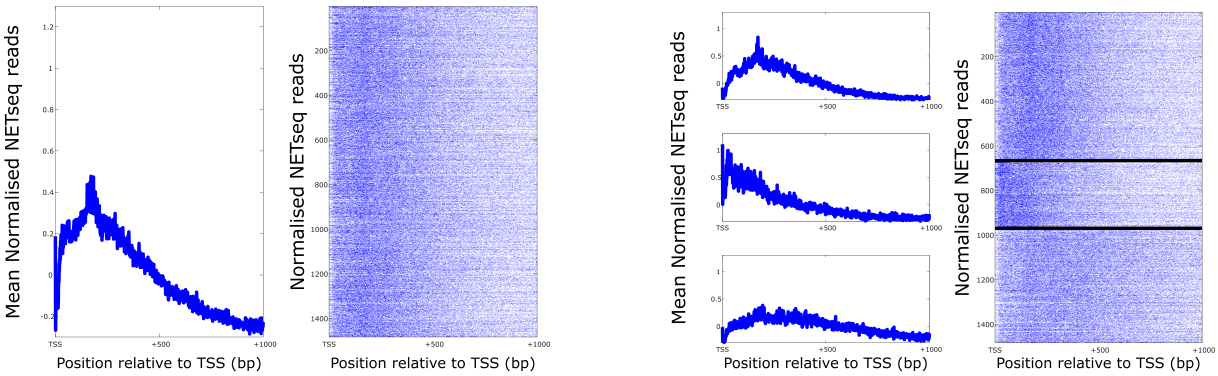


Figure 4.3: Representative profiles of NET-seq data from *S. cerevisiae*. Shown are the averaged profiles of all genes with signal above background (4.3a) and the averaged profiles of 3 groups clustered via K-means clustering (4.3b). NET-seq experiment was performed by Harry Fischl and is published in Fischl *et al.* 2017.

variation at different genes, moving beyond a view of a single factor performing a single role at every gene.

4.2 Polymerase profiles are distinct across the genome

The traditional method of displaying genome-wide datasets such as ChIP-seq or PRO-seq has been to average the signal of all - or a chosen subset of - genes, creating a metagene profile. This gives an indication of where the signal is most present relative to a particular feature, such as the transcription start site (See Figure 4.3a). While plotting data as an averaged metagene or a heatmap of the signal of each feature offers an immediate method of determining properties such as the distribution of signal, or even changes in levels between experiments, these methods present the opportunity to mask information which may be of real interest. Furthermore, the method by which metagenes are made has a large impact on the final profile obtained, and thus the conclusions one can draw from the data. Without being careful about the annotations of the features to be averaged over, determination of genes with extreme levels of signal or that sub-populations of genes with distinct profiles may exist, this method of bioinformatic analysis creates more issues than it claims to solve. Indeed, even true "genome-wide" techniques such as heatmaps (e.g. Figure 4.3a) are problematic as the eye is immediately drawn to any immediate pattern, which is sensitive to ordering of the data.

In an attempt to gain further information from the nascent transcription data col-

lected, we turned to clustering the profiles at each gene. Performing semi-supervised clustering via K-means clustering (Macqueen 1967), the NET-seq data from *S. cerevisiae* revealed at least three distinct groups of polymerase profiles (Figure 4.3b). Numbers of clusters were chosen such that each cluster exhibited a roughly distinct average profile and each cluster was of a similar size. This can be thought of as analogous to the "Elbow method" of clustering where the number of clusters is defined by the point where adding more clusters does not increase the explained variance by the model to a large degree. Clustering via hierarchical methods resulted in the identification of outliers as the most pronounced clusters, so was deemed inappropriate and PCA clustering could not be performed easily by eye and was certainly no better than k-means clustering. These groups showed (i) a broad peak of polymerase over the first 400bp of transcription units, typical of metagenes of this type, (ii) a group with an immediate peak of polymerase that dropped off by ∼250 bp after the TSS and (iii) a much flatter profile, with the polymerase spread fairly evenly over the body of the genes. The lists of genes in each clusters could not be ascribed to distinct gene ontologies, suggesting perhaps methods of expression and polymerase distinct from the fate of the transcript, potentially due to surrounding chromatin environment or location in the nucleus.

We were also able to confirm this result by performing the same downstream analysis to published *S. cerevisiae* NET-seq data from the Weissman lab. The original *S. cerevisiae* NET-seq data published in Churchman & Weissman 2011 differ from the data from Fischl *et al.* 2017, as it does not including a control step for the incorporation of RNA not part of the polymerase complex via a no-tag sequencing control. While the metagene profiles from the two studies differ slightly (Figures 4.3a & 4.4a), they still both display the characteristic pronounced 5' peak of polymerase, which diminishes throughout the gene body. Furthermore, after performing semi-supervised k-means clustering on the NET-seq data from Churchman & Weissman 2011, we still detect the three classes of genes as previously, with one cluster showing a broad peak of polymerase at the 5' end of the gene, one with a sharp peak immediately downstream of the TSS and a cluster with a flat, even distribution across the gene body (Figure 4.4b). The confirmation of these profiles of transcriptionally-engaged polymerase through the datasets from two independent labs

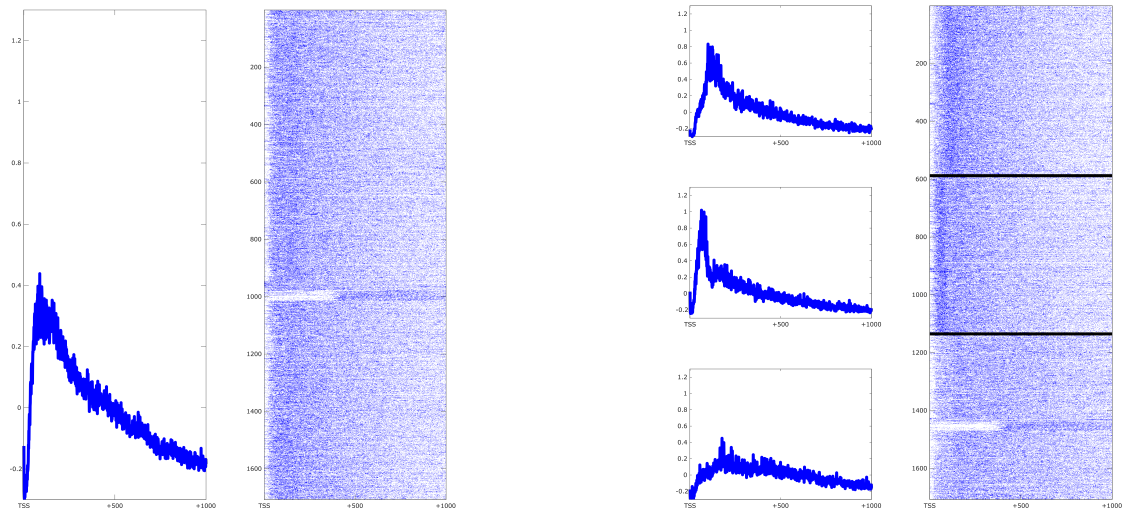


Figure 4.4: Representative profiles of NET-seq data from *S. cerevisiae*. Shown are the averaged profiles of all genes with signal above background (4.4a) and the averaged profiles of 3 groups clustered via K-means clustering (4.4b). For full axis labels see Figure 4.3. NET-seq data is taken from Churchman & Weissman 2011.

confirm the reproducibility of these distinct profiles; therefore, this begs the question, what causes these differing profiles of polymerase and are these differences functional?

There are a whole host of different distributions of polymerase that are seen at actively transcribed genes across the yeast genome. Of interest, therefore, is how these different profiles arise and what effects these have on gene regulation. It is clear that the polymerase undergoes a number of complex dynamic transitions throughout transcription and this heterogeneous pattern of polymerase distributions begs the question as to how these transitions vary across the genome and how individual regulatory factors influence transcription.

4.3 Polymerase profiles are independent of expression levels

One question it is not possible to determine the answer to using purely static steady-state data is whether the distributions of transcriptionally-engaged polymerase are determined by the level of transcriptional activity at each gene. Even using inducible systems such as

the glucose to galactose shift induces a degree of stress-response into the system, whereby the polymerase appears to occupy a more 5' position, independent of the resulting levels of mature RNA (Fischl *et al.* 2017). Therefore in order to look at how gene activation and repression affects the dynamics of polymerase in the gene body, we require a system whereby genes will undergo natural activation and repression and the dynamics of transcription elongation remain consistent to cells not undergoing dramatic stress-responses, such as is the case in the glucose-to-galactose shift (Fischl *et al.* 2017).

A useful system to study such a phenomenon is the Yeast Metabolic Cycle (YMC), whereby a population of cells synchronise their metabolic states and then proceed through the three main metabolic phases: a phase where yeast undergo high rates of oxygen consumption, termbed the Oxidative phase (OX), and two phases where cells cease oxygen consumption, with the first phase, termed Reductive Building (RB), occuring as dissolved oxygen levels in the media rise and the second, Reductive Charging (RC), covering the period of constant dissolved oxygen in the media before the yeast cells begin consuming oxygen again (See Figure 4.5, Tu *et al.* 2005, Kuang *et al.* 2014). The metabolic state of a cell is reflected in the amount of oxygen consumed by a cell, but also in the expression of roughly 60% of genes, which undergo oscillations in expression with the same period as the YMC. Now using the population of cells with synchronised metabolism, we can treat this population as a pseudo-single-cell and NET-seq experiments performed across a cycle will indicate the position of polymerase across genes as the total levels of polymerase and RNA at a given gene changes.

Clustering genes by their NET-seq profiles across 12 time-points taken from the YMC, we are able to distinguish genes that exhibit their peaks and troughs of expression at different times as well as group those genes which display distinct polymerase profiles (Figure 4.5). Once clustered, genes fall into distinct classes with expression peaking in the RB phase (e.g. cluster 4) or the OX phase (cluster 6) or genes which do not show much in the way of oscillatory gene expression behaviour (cluster 2). By looking at the distributions of polymerase normalised by total reads per gene, we observe that genes fall into these clusters not only by their total levels of transcription, but by the distributions of polymerase across the gene.

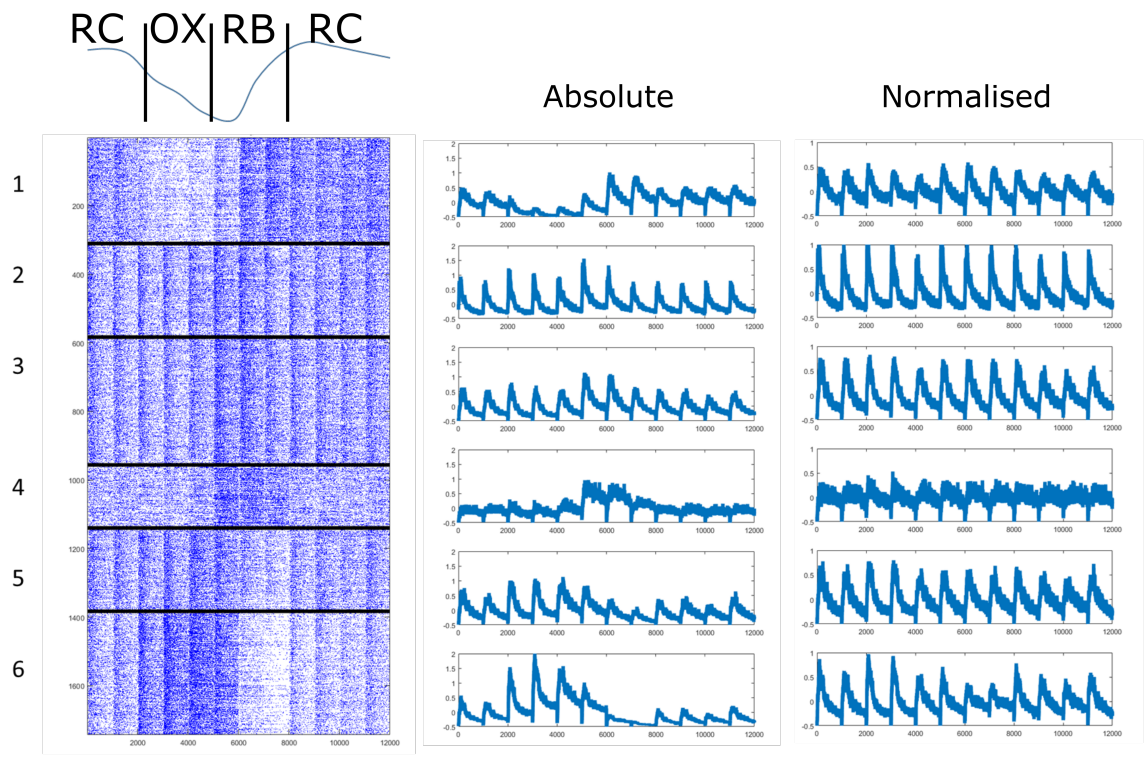


Figure 4.5: NET-seq distributions of genes through the yeast metabolic cycle. NET-seq data collected from cells undergoing synchronised oscillating metabolism. Shown are the data clustered by their NET-seq profiles across all 12 profiles (left), averaged NET-seq distributions for each cluster (middle) and metagenes repeated with total reads for each gene normalised to 1, then NET-seq signal averaged across each cluster (right). Experiments performed by Shidong Xi.

We observe genes which peak in their expression during the RB phase of the YMC, yet have a consistently flat distribution of polymerase across the gene regardless of expression level (cluster 4). Furthermore we observe genes which peak in the RB and through the RC phase (cluster 1) which have a distinct 5' bias in their polymerase distribution, regardless of total expression levels.

Furthermore, using the steady-state experiments to determine the levels of RNA at each gene, we also observe no correlation between the amount of RNA produced and the proportion of polymerase in the promoter-proximal region (Figure 4.6). The three groups of genes determined from clustering the genes as in Figure 4.3b also show no overall difference in the expression of genes in each cluster.

This points to a mode of transcription whereby the total level of polymerase transcriptionally engaged at a gene is determined by upstream factors such as transcription factors or regulatory networks. Once engaged with the DNA, the polymerase will un-

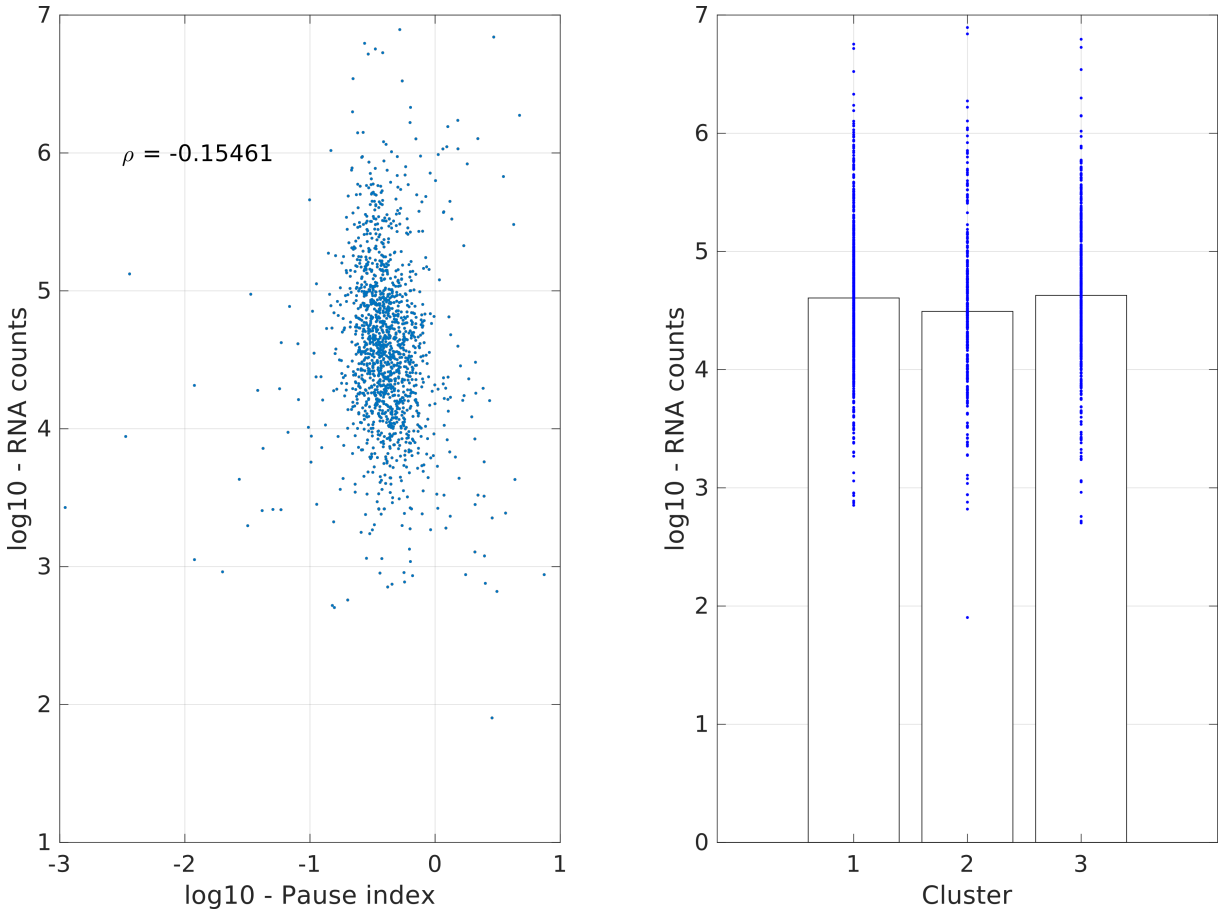
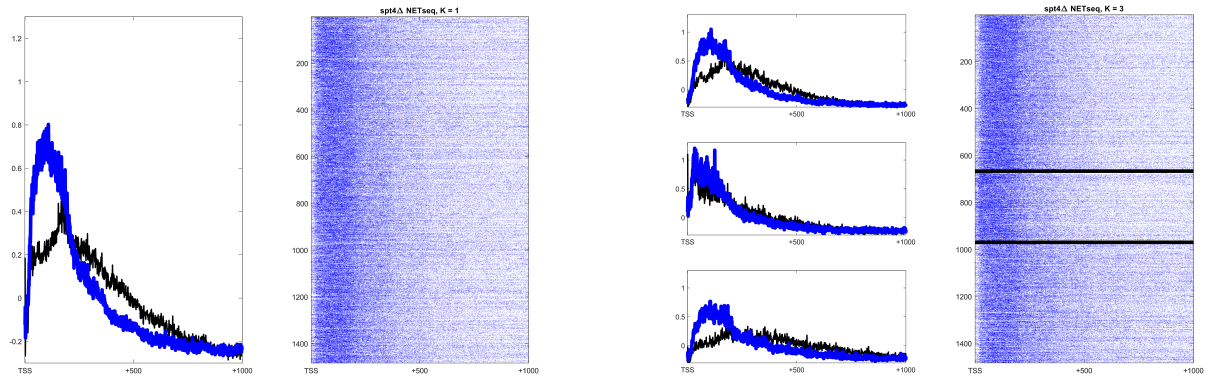


Figure 4.6: There is no correlation between polymerase distribution and gene expression. Shown (left) is the total number of RNA-seq counts from *S. cerevisiae* cells for each gene against the "pause-index", defined as the ratio of NET-seq signal in the first 200bp after the TSS to those from 201bp to 1,000bp downstream of the TSS. Correlation coefficient is Spearman's ρ. The levels of RNA-seq counts per gene are not significantly altered between clustered profiles (right). Each gene is shown with the mean level for each cluster given by bar height. Clusters are the same as those in Figure 4.3b. RNA-seq performed by Harry Fischl is the glucose-media dataset taken from Chapter 2.

dergo the same transitional steps at the same rate, regardless of the level of activity at the gene, suggesting that the dynamics of the polymerase through the gene is inherent to the gene regardless of levels of initiation. It appears that dynamic gene expression in this native system can be thought of more as being controlled by a "rheostat"-type mechanism, whereby the level of activity at the promoter can be switched from almost nothing through a continuum to very high levels. This is certainly in accordance with the idea of a promoter which allows a full dynamic range of activity, not simply one with on or off states (Corrigan *et al.* 2016).



(a) **(b)**

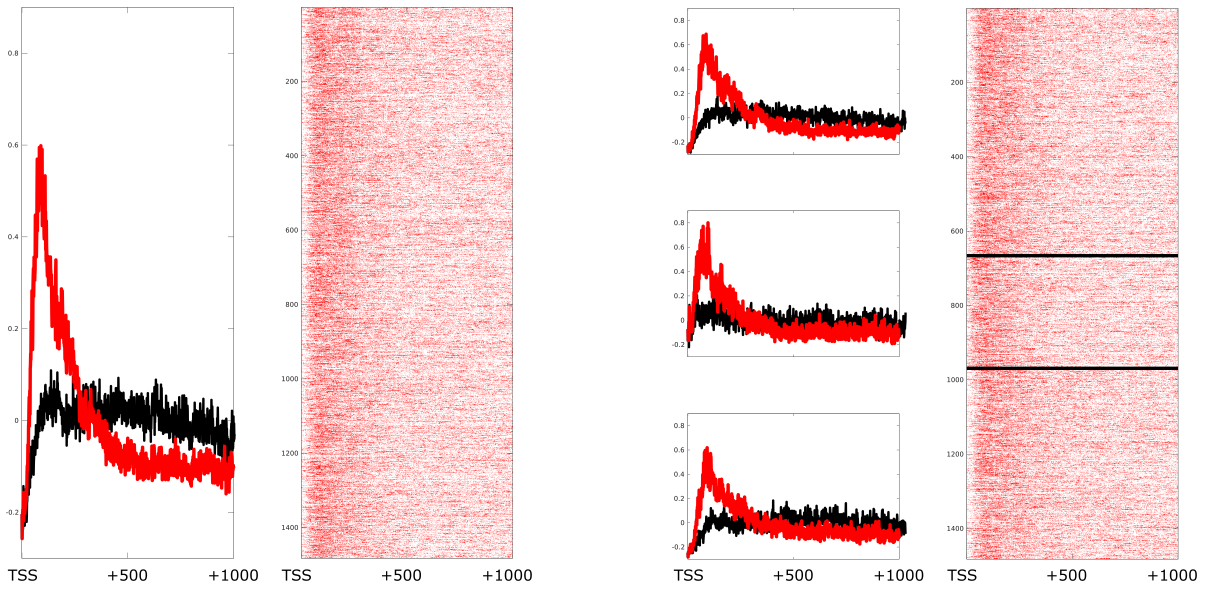
Figure 4.7: Representative profiles of NET-seq data from *S. cerevisiae spt4*Δ. Shown are the averaged profiles of all genes with signal above background (4.7a) and the averaged profiles of the 3 groups clustered via K-means clustering of NET-seq data (4.7b). The clustered gene groups are the same as those in Figure 4.3. Wild-type profiles are shown in black underneath. NET-seq experiments were performed by Ülkü Uzun and taken from Fischl *et al.* 2017.

4.4 Inhibiting elongation factors alter polymerase dynamics

4.4.1 Spt4 assists polymerase through early transcription

The typical *S. cerevisiae* NET-seq profile shows that the majority of polymerase is found immediately downstream of the TSS, with signal generally falling through the gene body (Churchman & Weissman 2011, Fischl *et al.* 2017). We then asked what role elongation factors have on creating this distribution of polymerase, and how these factors affect the distributions of transcribing polymerase. First, we asked how deletion of DSIF component Spt4 alters the distribution of polymerase as determined by NET-seq. Looking at an averaged metagene profile, deletion of the gene *SPT4* sees a much more pronounced peak of polymerase at the 5' end of genes in *S. cerevisiae* (Figure 4.7).

The plots shown here are normalised levels of polymerase across the genome, therefore we cannot immediately determine whether the distribution compared to the wild-type profiles (Figure 4.3) are made from more polymerase at the 5' end of genes, or less in the gene body. However, a comparative experiment published using PRO-seq (Booth *et al.* 2016 and Figure 4.8) analysing elongating polymerase in wild-type and *spt4*Δ cells with spike-in normalisation suggests that the reads in the gene-body are fairly similar between the two strains, with much more polymerase immediately downstream of the TSS.

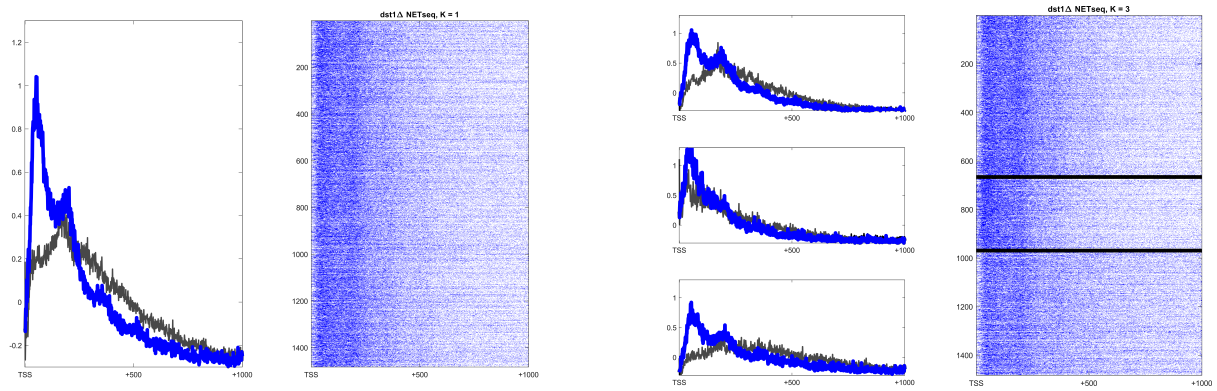


(a) (b)

Figure 4.8: Representative profiles of PRO-seq data from wild-type (black) and *S. cerevisiae spt4*Δ cells (red). Shown are the averaged profiles of all genes with signal above background (4.8a) and the averaged profiles of the 3 groups clustered via K-means clustering of NET-seq data (4.8b). The clustered gene groups are the same as those in Figure 4.3. For full axis labels see Figure 4.3.

Qualitatively, this suggest a role for Spt4 in the release of polymerase from the start of genes. Given the reported roles for DSIF in proximal-pausing of polymerase in metazoa (Kim & Sharp 2001), we may conclude that in the presence of Spt4, transcription in *S. cerevisiae* proceeds efficiently through the early gene template, potentially through dynamic transitions of the elongation complex.

Furthermore, for those genes in Cluster 2, which already display a pronounced peak of polymerase at the 5' end of the gene, there appears to be no great change in the distributions of polymerase upon deletion of Spt4. This suggests either that the elongation complex at these genes may not include Spt4, resulting in a build-up of polymerase in the promoter-proximal region or that these genes have reached a saturation point of promoter-proximal polymerase. Indeed, as noted in Ehrensberger *et al.* 2013 and Battaglia *et al.* 2017, there should be a theoretical limit to the rate at which polymerases that can occupy the promoter regions of genes which is defined by the size of the polymerase footprint and the time it takes for the polymerase to move enough nucleotides forward to allow another polymerase to bind.



(a) **(b)**

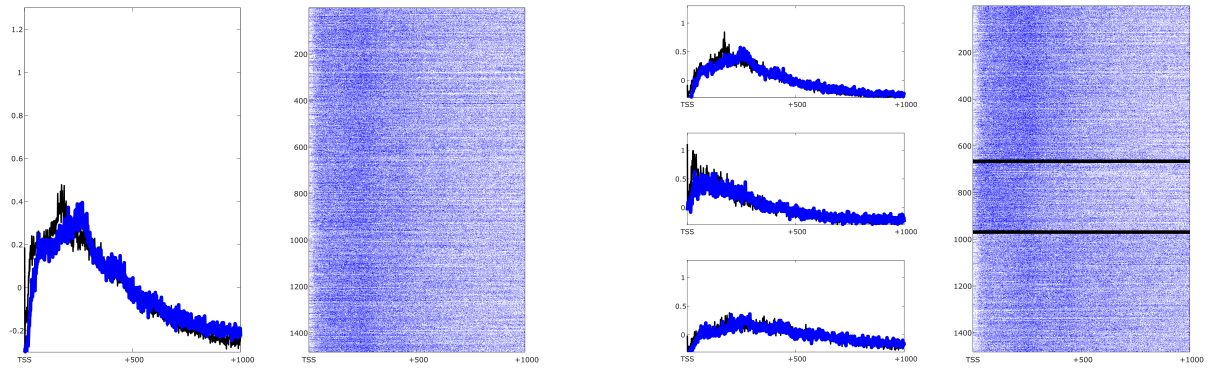
Figure 4.9: Representative profiles of NET-seq data from *S. cerevisiae dst1*Δ. Shown are the averaged profiles of all genes with signal above background (4.9a) and the averaged profiles of the 3 groups clustered via K-means clustering of NET-seq data (4.9b). The clustered gene groups are the same as those in Figure 4.3. Wild-type profiles are shown in black underneath. For full axis labels see Figure 4.3. NET-seq experiments were performed by Ülkü Uzun and taken from Fischl *et al.* 2017.

4.4.2 Backtrack release determines polymerase location

The second genetic intervention made was to delete the backtrack release factor TFIIS, Dst1 in *S. cerevisiae*. It has been previously reported (Churchman & Weissman 2011) that deletion of *DST1* results in an altered polymerase profile. Consistent with that study, we saw a much more pronounced 5' peak, similar to that seen under deletion of *SPT4* (Figure 4.9). With this more pronounced peak in polymerase, there are also more positional build-ups of polymerase at certain locations through the gene body (see Figure 4.9b). This suggests polymerase enters into a number of backtracked states at pre-defined locations throughout the gene body while transcribing, which must then be recovered by Dst1 activity.

4.4.3 What role does Paf1 have in regulating early transcription?

As was noted in Fischl *et al.* 2017, the NET-seq profiles from a *paf1*Δ show a similar profile to wild-type cells, but with a downstream shift in the location of the main peak of polymerase accumulation (Figure 4.10a). By looking at the profiles of sub-clusters of genes, we observe a loss of any accumulation of polymerase immediately downstream of the TSS in the second cluster defined in Figure 4.3b, which has the most pronounced



(a) **(b)**

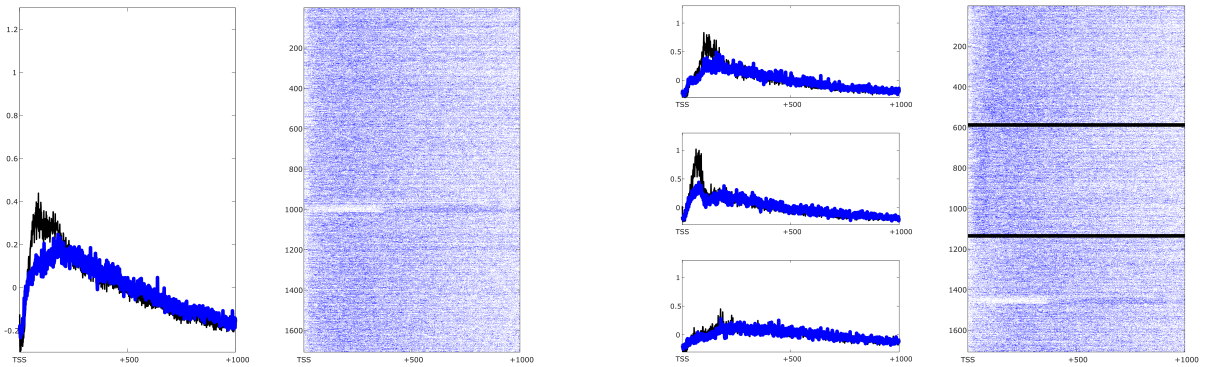
Figure 4.10: Representative profiles of NET-seq data from *S. cerevisiae paf1*Δ. Shown are the averaged profiles of all genes with signal above background (4.10a) and the averaged profiles of the 3 groups clustered via K-means clustering of NET-seq data (4.10b). The clustered gene groups are the same as those in Figure 4.3. Wild-type profiles are shown in black underneath. For full axis labels see Figure 4.3. NET-seq experiments taken from Fischl *et al.* 2017.

5' peak of polymerase (Figure 4.10b), however there does not appear to be a consistent genome-wide change to these profiles. This observation of more polymerase in the body of the gene relative to the promoter-proximal region at a subset of genes is in accordance with observations made in Chen *et al.* 2013 that PAF1 may act as a protein to maintain the paused-state of the polymerase after initiation in human cells. However, it is in contrast to results from Yu *et al.* 2015 that PAF1 acts to release from the promoter-proximal pause. Both of these conclusions were made from qualitative observations made from run-on and ChIP-seq experiments with low resolution and, in the case of ChIP-seq, no strand-specificity. While the observations made here do not offer an immediate genome-wide explanation for the role of Paf1 in *S. cerevisiae* transcription, quantitative approaches presented in the next Chapter will elucidate the matter further.

4.5 How do histone writers influence distributions of polymerase?

4.5.1 The role of histone acetylation on transcription dynamics

Two components of the histone de-acetylase complex Rpd3S were investigated in Churchman & Weissman 2011, with NET-seq experiments performed in strains lacking the sub-units Eaf3 and Rco1. Although the two components form an interacting sub-unit of



(a) **(b)**

Figure 4.11: Representative profiles of NET-seq data from *S. cerevisiae eaf3*Δ. Shown are the averaged profiles of all genes with signal above background (4.11a) and the averaged profiles of the 3 groups clustered via K-means clustering of NET-seq data (4.11b). The clustered gene groups are the same as those in Figure 4.4. Wild-type profiles are shown in black underneath. For full axis labels see Figure 4.3. NET-seq data taken from Churchman & Weissman 2011.

this Histone De-ACetylase (HDAC), Eaf3 is also a component of the Histone Acetyl-Transferase (HAT) NuA4. As such, although we may expect similar outcomes from deletion of the two regarding their de-acetylation activity, the effects of Eaf3 on transcription may be entirely through its role in acetylating histones, or may be the result of two conflicting actions on the chromatin. Deletion of *EAF3* results in a less well-pronounced peak of polymerase at the 5' region of genes in the two clusters of genes with 5' biases of polymerase (Clusters 1 & 2, Figure 4.11b). However, little change is observed in the distributions of polymerase in those genes which do not display 5' peaks of polymerase in wild-type cells (Cluster 3, Figure 4.11b).

This suggests a role for Eaf3 in maintaining polymerase in the promoter-proximal region, supposedly through either the placement or removal of acetyl marks on the histones. However, this effect is not seen at all genes and the activity of Eaf3 may be restricted to certain genes. This would certainly agree with conclusions made in Brown *et al.* 2018 that Set3 HDAC activity is less pronounced on genes with antisense transcription, where antisense transcription acts to buffer acetylation loss at these genes. There may be other unknown components which act with either Rpd3S or NuA4 and Eaf3 to maintain "normal" transcription dynamics at the sub-class of genes without proximal peaks of polymerase or altered polymerase distributions after deletion of *EAF3*.

In contrast to Eaf3, removal of Rco1 does not appear to affect polymerase distributions at transcribed genes (Figure 4.12). Any changes at any of the clusters of polymerase

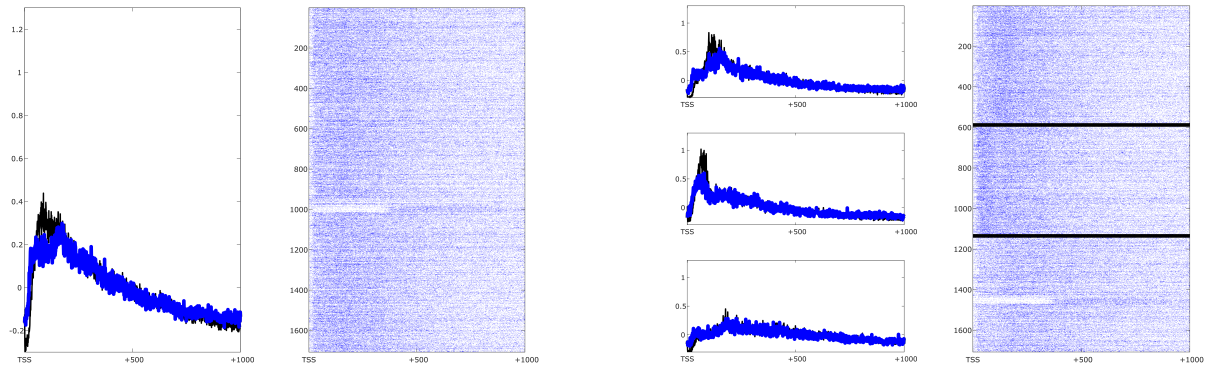
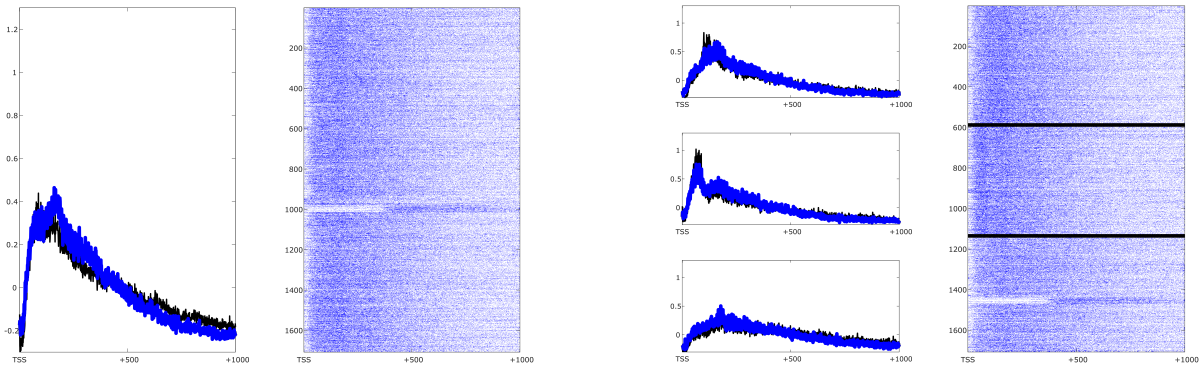


Figure 4.12: Representative profiles of NET-seq data from *S. cerevisiae rco1*Δ. Shown are the averaged profiles of all genes with signal above background (4.12a) and the averaged profiles of the 3 groups clustered via K-means clustering of NET-seq data (4.12b). The clustered gene groups are the same as those in Figure 4.4. Wild-type profiles are shown in black underneath. For full axis labels see Figure 4.3. NET-seq data taken from Churchman & Weissman 2011.

profiles are very marginal, suggesting that either Rco1 has a degree or redundancy in cells, or the Rpd3S complex has functional roles outside of regulating transcription dynamics. Churchman & Weissman 2011 noted that in the cells lacking either component of Rpd3S, bidirectional transcription was much more widespread out of promoters in the non-coding direction and given the role of Rpd3S in de-acetylating histones in the terminator region of long genes (Carrozza *et al.* 2005), rather than the gene body, this indicates that it is not only the level of histone aceylation which is important in governing transcription dynamics, but where the modified nucleosomes are in relation to the TSS.

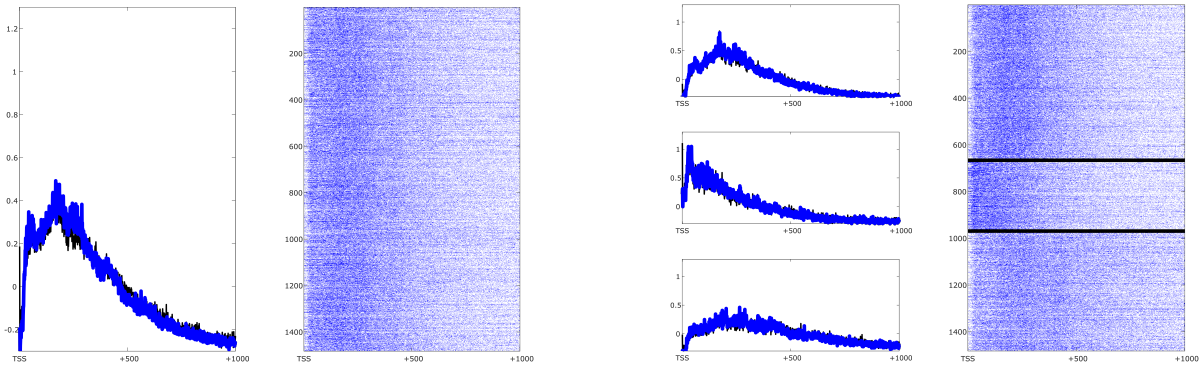
4.5.2 Histone methyltransferases do not have pronounced effects on sense transcription

Two further proteins investigated in Churchman & Weissman 2011 are the histone methyltransferases Set1 and Set2. Set1 is a subunit of the COMPASS (Set1C) complex, responsible for the methylation of H3K4, a mark of active transcription. We observe little difference in the profiles of transcriptionally-engaged polymerases upon loss of Set1, either at the genome-wide averaged levels or within the defined clusters of polymerase profiles (Figure 4.13). We also compared this with the NET-seq data taken from strains lacking *SPP1*, a further subunit of the COMPASS complex, this time responsible for the tri-methylation of H3K4. As H3K4me3 is particularly associated with highly transcribed



(a) **(b)**

Figure 4.13: Representative profiles of NET-seq data from *S. cerevisiae set1*Δ. Shown are the averaged profiles of all genes with signal above background (4.13a) and the averaged profiles of the 3 groups clustered via K-means clustering of NET-seq data (4.13b). The clustered gene groups are the same as those in Figure 4.4. Wild-type profiles are shown in black underneath. For full axis labels see Figure 4.3. NET-seq data taken from Churchman & Weissman 2011.

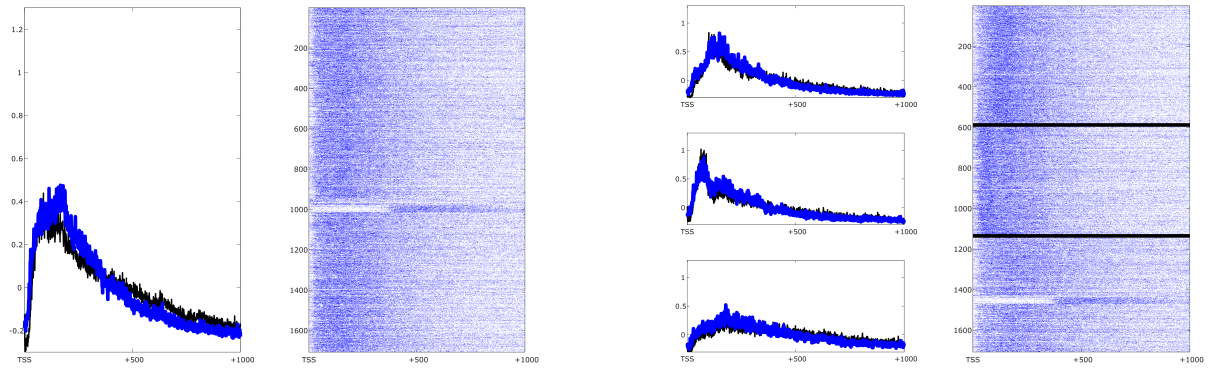


(a) **(b)**

Figure 4.14: Representative profiles of NET-seq data from *S. cerevisiae spp1*Δ. Shown are the averaged profiles of all genes with signal above background (4.14a) and the averaged profiles of the 3 groups clustered via K-means clustering of NET-seq data (4.14b). The clustered gene groups are the same as those in Figure 4.3. Wild-type profiles are shown in black underneath. For full axis labels see Figure 4.3. NET-seq data taken from Howe *et al.* 2017.

genes, we are interested in whether the histone mark is at all instructive for transcription or the dynamics of polymerase. We again observe no effect on the profiles of transcribing polymerase after removal of Spp1, suggesting that the modifications of this histone residue do not influence the elongation dynamics (Figure 4.14).

It has been recently suggested that the methylation state of H3K4 of nucleosomes in genic regions may be a by-product of transcription. It has been reported that during the YMC with oscillating expression of genes, RNA appears at genes before H3K4me3 (Kuang *et al.* 2014, Howe *et al.* 2017) and that if transcription is induced, the length of time a gene is transcribed dictates the methylation state of H3K4, with the three methyl groups



(a) **(b)**

Figure 4.15: Representative profiles of NET-seq data from *S. cerevisiae set2*Δ. Shown are the averaged profiles of all genes with signal above background (4.15a) and the averaged profiles of the 3 groups clustered via K-means clustering of NET-seq data (4.15b). The clustered gene groups are the same as those in Figure 4.4. Wild-type profiles are shown in black underneath. For full axis labels see Figure 4.3. NET-seq data taken from Churchman & Weissman 2011.

added as transcription proceeds over time (Soares *et al.* 2017). The question then arises, if these marks are not instructive for transcription or gene expression, what purpose do they serve in the cell? The methylation of H3K4 has been associated with the recruitment of chromatin remodelling factors, which may be brought to the chromatin to 'clear up' after the polymerase has passed through, allowing further rounds of transcription (Santos-rosa *et al.* 2003). It has been recently suggested that Spp1 may act as a cell-cycle modulated protein, rather than via transcription, and is involved in the formation of a chromatin structure in the preparation of cell division (Karányi *et al.* 2018).

The second histone methyltransferase investigated in Churchman & Weissman 2011 was Set2, which methylates the H3K36 residue (Strahl *et al.* 2002) and is responsible for the further recruitment of the histone de-acetylatse complex Rpd3S (Keogh *et al.* 2005). Similar to the loss of the H3K4 methyltransferases, the loss of Set2 did not result in a large change in the distributions of transcribing polymerase (Figure 4.15). This once again shows that changing the methylation state of the underlying chromatin does not have a consistent effect on the dynamics of transcription or the elongation process.

4.6 Polymerase profiles in mammalian cells

A tandem technique to NET-seq has been developed in a *H. sapiens* cell line, termed mNET-seq (mammalian NET-seq, Nojima *et al.* 2015), again tracking the location of

transcriptionally-engaged polymerase with RNA in the active site. It has been reported that transcription in HeLa cells is regulated post-initiation by the Early Elongation Checkpoint (EEC), or promoter-proximal pause. As such, we may expect a much more defined peak of polymerase immediately downstream of the TSS by this technique.

While the human and yeast genomes exhibit similar degrees of crowding regarding the overlapping nature of transcription units in the two organisms, the scales are vastly different. While *S. cerevisiae* genes have lengths on the order of 1,000bp, with a maximum size of ∼15,000bp, human genes are on the order of 10,000bp, with some genes reaching well over 2,000,000bp. Previous studies have represented the polymerase profiles using length-scaled average plots, often constricting the promoter-proximal region to a very small window of an aggregate plot (e.g. Nojima *et al.* 2015). Here, to make direct comparisons between organisms, we focus on the first 1,000bp of the transcription units. This allows us to directly compare the regulatory steps in transcription between *S. cerevisiae*, which supposedly lacks a promoter-proximal pause, and *H. sapiens*.

mNET-seq profiles from HeLa cells indeed show a much more pronounced peak in polymerase immediately downstream of the TSS than *S. cerevisiae* when viewed as a genome-wide average (Figure 4.16a). However, clustering genes by their profiles highlights the different profiles that exist, especially in relation to the location of the pronounced peak of polymerase (Figure 4.16b).

The alternative locations of polymerase peaks relative to the TSS potentially point to differential sites of promoter-proximal pause in these sub-classes of genes. One important aspect to highlight is that transcription cannot be thought of a universal process which is identical at all transcription units, in particular transcription of protein-coding genes is not a single process for every gene. Of particular interest is the class of genes that fall into the first cluster above (Figure 4.16b), which appear to have no build-up of polymerase in the promoter-proximal region, yet have mNET-seq signal above background, therefore are considered "expressing".

One immediate question that arises from looking at polymerase profiles is the effect on the downstream functional molecules produced such as mRNA and protein. To start investigating this question, we took a published RNA-seq dataset in HeLa cells from Mayer

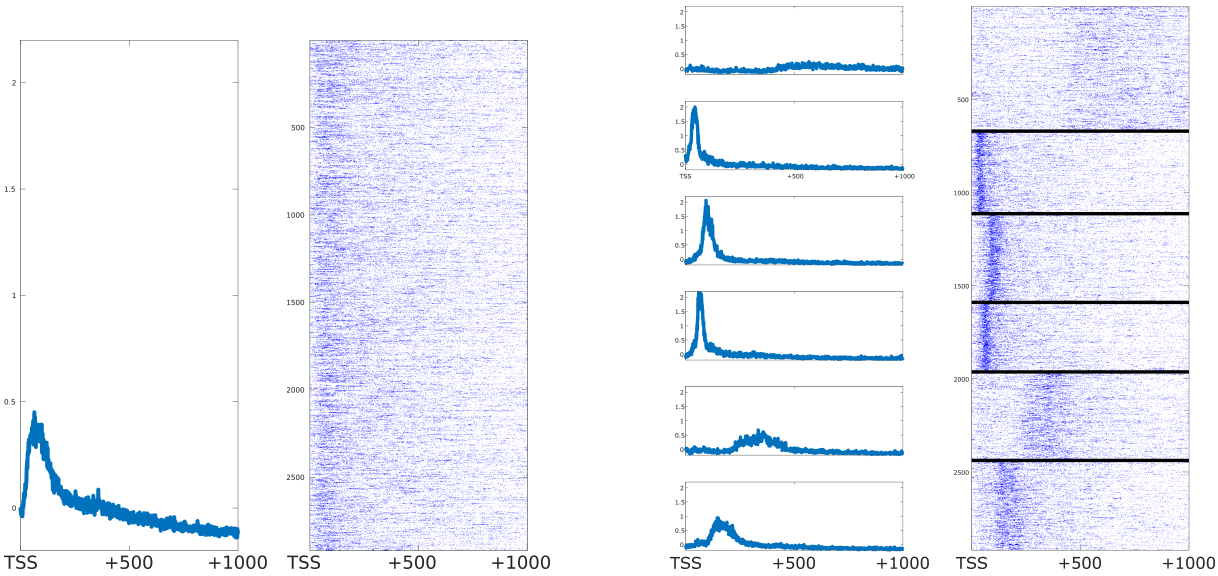


Figure 4.16: Representative profiles of mNET-seq data from HeLa cells. Shown are the averaged profiles of all genes with signal above background (4.16a) and the averaged profiles of 5 groups clustered via K-means clustering (4.16b). For full axis labels see Figure 4.3. mNET-seq experiments performed by Michael Tellier.

et al. 2015, comparing the number of RNA-seq reads at each gene to the polymerase profiles. Comparing number of RNA-seq reads to the pause index defined by mNET-seq showed no correlation (Figure 4.17), suggesting that degrees of proximal-pausing or polymerase processivity do not greatly impact the amount of transcript produced at each transcription unit. Furthermore, taking the clusters of genes as defined by their polymerase profiles as in Figure 4.16b showed no great difference in RNA levels (Figure 4.17). This is in accordance with the observations made in *S. cerevisiae* in batch culture and through the YMC, that overall expression of a gene does not seem to influence the distribution of polymerase across a gene. This points to a role of having the majority of polymerase at the 5' end of the gene beyond merely "priming" a repressed gene for activation upon an external cue (Rougvie & Lis 1988).

While the transcription dynamics leading to these distinct polymerase profiles are complex and point to different mechanisms of regulation, the fact that downstream levels are not affected suggests further mechanisms buffering any differences at the level of transcription. It may, however, be the case that the chromatin template the polymerases must pass through are affected by other elements, such as enhancers or 3-dimensional chromosome conformation and vice-versa.

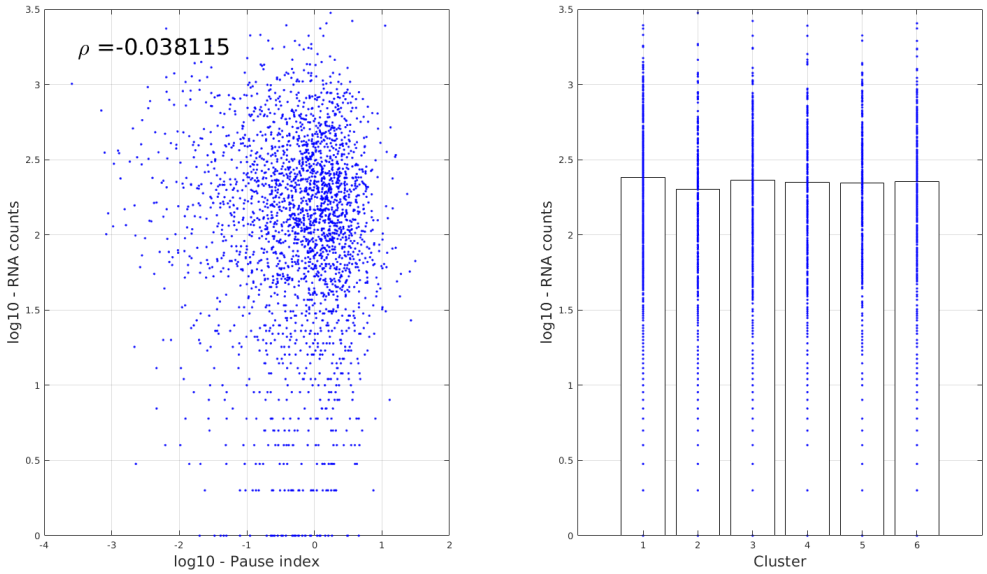


Figure 4.17: There is no correlation between polymerase distribution and gene expression in HeLa cells. Shown (left) is the total number of RNA-seq counts from HeLa cells for each gene against the "pause-metric", defined as the ratio of mNET-seq signal in the first 200bp after the TSS to those from 201bp to 1000bp downstream of the TSS. Correlation coefficient is Spearman's ρ. The levels of RNA-seq counts per gene are not significantly altered between clustered profiles (right). Each gene is shown with the mean level for each cluster given by bar height. Clusters are the same as those in Figure 4.16b. HeLa mNET-seq experiments performed by Michael Tellier from the lab of Shona Murphy, RNA-seq taken from Mayer *et al.* 2015.

4.6.1 Post-translational modifications of elongation factors determine polymerase profiles

The genetics of *S. cerevisiae* allows for deletion of entire regulatory factors and the observation of downstream effects of the loss these proteins. Studies in human cells have mostly relied on depletion of factors through the RNAi machinery, using specific drugs to target the catalytic effects of proteins or using the recently developed CrispR system to modify elements of the genome (Zheng *et al.* 2014).

One drug which is well annotated as having a pronounced effect on transcription elongation is DRB, from which the alternative name of the SPT4/5 complex DSIF arises (DRB-Sensitivity Inducing Factor). DRB has been shown to target the SPT4/5 complex by inhibiting the phosphorylating action of CDK9 (also known as Positive Transcription Elongation Factor b, P-TEFb) (Laitem *et al.* 2015).

Here, we wish to investigate what the direct effects are on transcription elongation after inhibition of CDK9, how quickly these effects arise and whether these are universal across

the genome. We analysed mNET-seq data in HeLa cells before and after treatment with DRB for 5, 10 & 15 minutes and observed a build-up of polymerase at the 5' end of genes which already exhibited an accumulation of polymerase (Figure 4.18). By performing k-means clusters on the mNET-seq profiles across all 4 timepoints, we found genes which exhibited even distributions of polymerase across the gene and 3 clusters with increasing levels of polymerase in the promoter-proximal region. Interestingly, we found that at those genes which did not exhibit a peak of polymerase at the 5' region of the gene that there was no obvious change to the distribution of polymerase and certainly not to the same degree as the other clusters of genes.

The phosphorylation of SPT5 during the transitions from transcription to processive elongation has already been suggested to control the duration of promoter-proximal pauses in higher eukaryotes (Kwak & Lis 2013). Here, we would suggest that CDK9 only acts on a subset of genes, causing polymerase to accumulate immediately downstream of the TSS and loss of the phosphorylation of SPT5 results in a larger build-up of polymerase in this region. The established model of the transitions of polymerase through these stages would then suggest that NELF is still engaged with the polymerase complex, resulting in a stable paused polymerase waiting for its cue to proceed into elongation.

That loss of CDK9 has an effect on transcription elongation within 5 minutes shows how dynamic the process of transcription is and how these transitions are occurring on a short time-scale. We are further interested in what the effects are on the rates of polymerase elongation, stalling and early termination after CDK9 inhibition and will rely on more sophisticated modelling techniques to unpick this, which will be discussed in Chapter 5.

4.7 Discussion

The dynamics of transcription are clearly complex, with the polymerase complex undergoing a number of transitions from initiation to termination and with a host of factors contributing to this process. In *S. cerevisiae*, there are distinct sub-groups of genes whose transcription dynamics are set apart from each other. We showed in Chapter 3 that both antisense transcription and the levels of histone acetylation can affect the rates

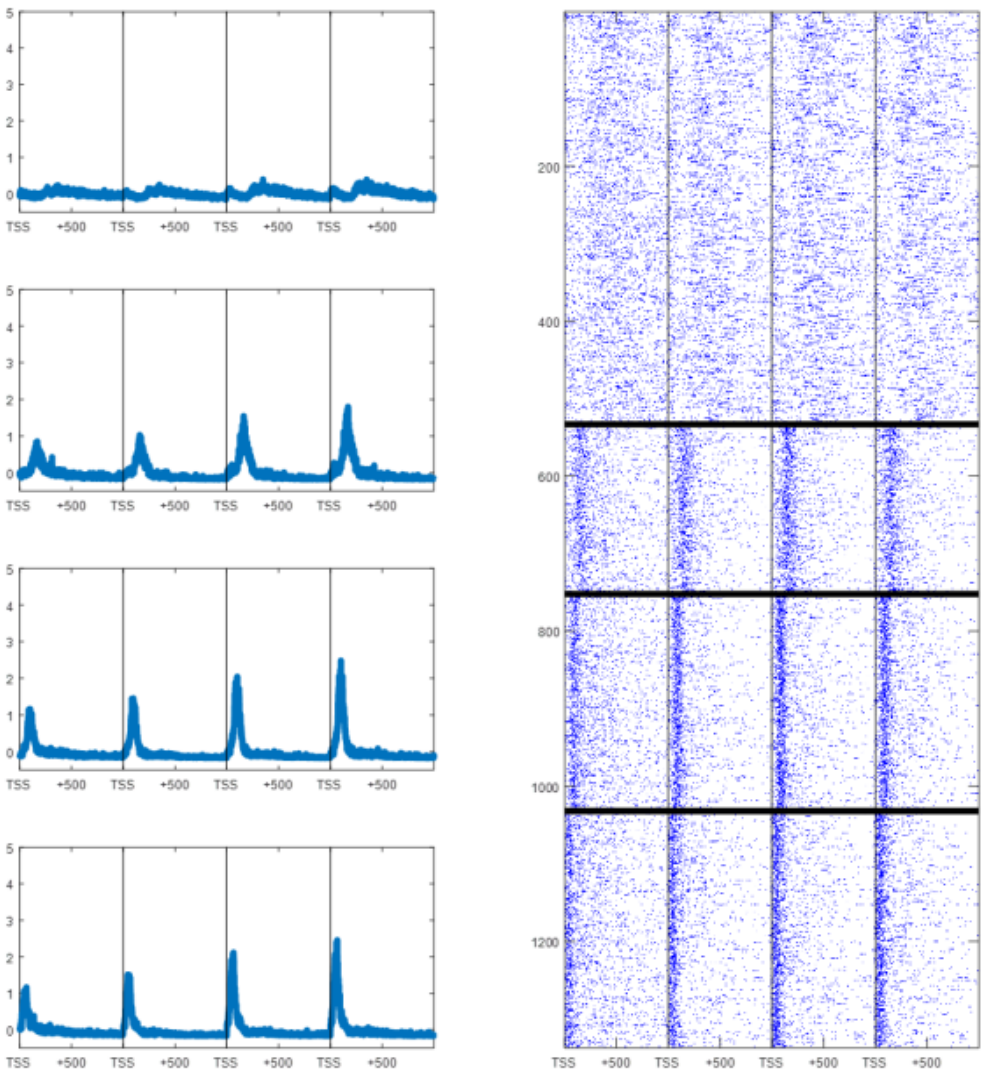


Figure 4.18: mNET-seq profiles in HeLa treated with DMSO and then with CDK9 inhibitor DRB for 5, 10 & 15 minutes. Genes were clustered into 4 groups via k-means clustering of the mNET-seq profiles across all 4 timepoints. Shown are absolute metagene profiles of each cluster (left) and heatmaps showing the signal at every gene (right). Genes were only included if they were found to be expressing above background in every timepoint.

of transcription initiation and elongation at an individual gene. Here, in moving to a genome-wide view of transcription, we encounter a wider range of transcription dynamics and a wide range of factors which act to different degrees at each gene. Furthermore, different elongation factors have been shown to preferentially bind to sub-sets of genes, potentially influencing both the transcription and processing of the RNA of a gene (Fischl *et al.* 2017, Battaglia *et al.* 2017).

Here, we identified at least 3 clusters of genes with distinct polymerase distributions in *S. cerevisiae*, set apart by the proportion of polymerase found in the promoter-proximal

region relative to the gene body. This separation was more pronounced in HeLa cells, with polymerase almost exclusively found near the promoter at some genes, with others showing even spreads of polymerase throughout the gene body. It is difficult to speculate what sets these genes apart, particularly without knowledge of the different obstacles the elongation complexes will face at these different genes, however this clearly points to alternative activity of the polymerase at these different groups of genes. We showed that elongation factors previously described as influencing the levels of pausing or backtracking in cells such as Spt4 and Dst1 have important roles in establishing the kinetics of RNA PolII transcription. Interfering with these polymerase-associated proteins in *S. cerevisiae* interrupt the progress of the elongating polymerase, causing polymerase to accumulate further in the promoter-proximal region of the genes, however the degree to which these profiles change varies between genes, with some genes more sensitive to loss of Spt4 and Dst1 than others. The interplay between these elongation factors are likely crucial in maintaining the correct balance of RNA and proteins in the cell, especially under stress. Previous studies have shown that removing or depleting elongation factors cause widespread changes to the dynamics of transcription in cells as well as the balance of RNA in a steady-state population (Fischl *et al.* 2017, Chen *et al.* 2015), suggesting that even though there is no apparent correlation between polymerase distributions on expression levels, disruption of the former still has the capacity to impact on the latter.

We further found that while post-translational modifications on histones have been associated with transcription and gene expression in the past, we do not see a profound effect on the distributions of polymerase after deleting the writers and erasers of the histone residues. Interestingly, the two subunits of the histone de-acetylase Rpd3S complex did not result in the same changes to the distributions of polymerase, however this is likely due to further roles of Eaf3 in the HAT NuA4. Deletion of Eaf3 in *S. cerevisiae* saw a loss of the 5' accumulation of polymerase at genes without flat RNA PolII profiles, but the same was not observed after deleting Rco1. This suggests a that although Eaf3 is well annotated as an interacting partner with Rco1 in the Rpd3S HDAC complex, the compensatory role of Eaf3 to acetylate histones as part of the NuA4 complex impacts the dynamics of polymerase elongation.

We also observed no great impact on the polymerase profiles after deletion of proteins responsible for methylating histone residues. As recent studies have shown, however, this may not be of surprise if the increased methylation of histones is a bi-product of transcription, impacting on other processes in the cell such as cell division (Howe *et al.* 2017, Soares *et al.* 2017, Karányi *et al.* 2018).

It would perhaps be advisable to investigate further the obstacles that the polymerase will face on their passage through the gene and an investigation of the known DNA-binding proteins that are found in intragenic regions may reveal which factors contribute to blocking polymerase passage, if there are any. The resolution required is sufficiently high that techniques such as ChIP-seq would not be of assistance in this scenario, it is more important to know exactly where the protein is binding, either through known DNA-binding motifs, or interrogation via high-resolution cross-linking techniques such as a DNA-selective crac-style experiment or ChIP-exo.

However, the DNA and associated proteins are by no means the only factors which affect polymerase elongation. Without a full understanding of every factor which binds to the elongation complex at every stage of the transcription cycle and how these factors interact with each other and with polymerase and affect its post-translational modifications, we are unable to pinpoint exactly which points are crucial to the functional progression of polymerase.

Finally, the analysis performed here has been merely qualitative in describing the changes in transcription entirely through the changes in distributions of transcriptionally-engaged RNA PolII. While this offers a start to understanding the roles of these factors in the process of transcription, we wish to understand in greater detail the dynamics of the transcription process and how these factors work together to establish the kinetics of polymerase seen in wild-type cells. In the next Chapter, we construct mathematical models of transcription and use them to determine dynamic rates for the key steps in polymerase elongation and in turn the roles of individual factors in establishing these rates.

Chapter 5

RNA polymerase II exists in a number of states

Simulation scripts for this Chapter are available at: `https://github.com/tbrown91/NETSeq_simulator`.

5.1 A stochastic model of polymerase elongation

In the previous Chapter, we outlined the types of distributions of transcriptionally-engaged polymerase that are seen in eukaryotic organisms using the most recent techniques to study nascent transcription. The conclusions we are able to draw from bioinformatic analysis are useful, but currently remain qualitative descriptions of the distributions of polymerase, with some informed speculation of how these distributions may arise. We wish to further investigate how the different distributions of polymerase are reflected in the underlying mechanics and dynamics of the elongation process.

After inhibition of various factors involved with transcription and chromatin, we showed that this can impact the distributions of actively transcribing polymerase across the *S. cerevisiae* and human genomes. We are able to begin to speculate how these changes in distributions might arise given previous work on individual factors, yet we are unable to determine to what degree the factors affect individual genes and particular steps of the elongation process. Furthermore, when investigating the effect of factors on transcription for which there is no previous knowledge, we are left in the dark about how

this is truly affecting transcription and can only rely on comparisons with other factors, which may be misleading.

In order to add a more rigorous approach to investigating both the dynamics of endogenous transcription *in vivo* and the effects of regulatory factors, we turn to mathematical modelling of the process of elongation as a tool to explain the distributions of polymerase as determined by NET-seq and mNET-seq. Here, we start with simple models of polymerase transcription incorporating initiation and elongation and add additional mechanisms of interest including polymerase stalling and backtracking, early termination and more complex promoter activity.

5.1.1 Aims

First, we use NET-seq data from wild-type *S. cerevisiae* cells to determine the best-fit model of transcription to use for modelling NET-seq data. As highlighted previously, the process of transcription is known to undergo a number of complex steps, therefore we must incorporate the necessary reactions from this list to be able to construct a meaningful model which reflects the underlying dynamics. We wish to avoid over-fitting the data, while still incorporating enough detail in the model to account for the complex steps involved in transcription. As such, we use Information Criteria to judge model fits to the data weighted against the number of free parameters, or the complexity of each model. We then use this best fit model to determine the ranges of transcription dynamics that exist across the yeast genome and how this is reflected in the NET-seq profiles.

Secondly, we investigate the changes in transcriptional parameters after performing genetic intervention. Modelling the NET-seq distributions in strains following deletion of elongation factors, mutating the polymerase molecule and interfering with the RNA processing pathway, we determine the effect these interventions have on the dynamics of nascent transcription and thus how these factors contribute to the dynamics of transcription in the wild-type scenario. Using this approach to analyse high-resolution quantitative data, we are able to obtain numerical estimates on key transcription features such as the extent of polymerase stalling or backtracking at each gene and the processivity of a gene, defining the efficacy of transcription events.

Finally, we repeat this approach to investigate the dynamics of transcription in a human cell line. By testing the fit of transcription models to mNET-seq data from HeLa cells, we are able to to determine to what extent the dynamics from *S. cerevisiae* overlap with those in human cells. We then use this approach to analyse the effect of inhibiting the kinase CDK9 on the dynamics of transcription after rapid inhibition through treatment.

The study of the effects of these individual factors on transcription will feed together to inform the full picture of how the dynamics of transcription elongation are established, with many factors found to act together and antagonistically to varying degrees across eukaryotic genomes.

5.2 Models of polymerase elongation

To construct a model of polymerase elongation, we turned to existing models and looked to incorporate further dynamic steps into the model to capture the more complex nature of polymerase (Choubey *et al.* 2015, Sigurdsson *et al.* 2010, Larson *et al.* 2014). We chose to model the inherently noisy process of transcription via a stochastic model, building on the deterministic models established in other labs (Ehrensberger *et al.* 2013, Wu *et al.* 2016, Azofeifa & Dowell 2017). The aims of this section are twofold: (i) to determine the best model that explains the data and (ii) to find the rates of the steps in the model that create the profiles and in particular how these rates change upon intervention.

As the true structure of the underlying mechanisms is assumed to be complex, with the polymerase known to transition between elongating, stalled and backtracked states, we wish to find the combination of these reactions that best explains the data, without over-fitting. To determine the best model to explain the data, we simulated transcription, then took the final polymerase distribution as the total number of counts at each nucleotide of 10,000 cells. For ease of comparison the simulated gene was set at 1,000bp and compared to the first 1,000bp of all genes with annotated genes greater that 1kb in length not overlapping other transcription units.

5.2.1 Rules of the models of transcription

The models presented here are intended to include as many known steps of the transcription process as possible and we aim to find the model, or set of models, that result in a best-fit to the raw data without over-fitting. Only by testing models with a high number of unknown parameters will we be able to ascertain the point of over-fitting. Furthermore, a number of the models are inclusive of each other and the simpler models are obtainable from the more complex models by setting the rates of the extra reactions to zero. This would immediately identify a redundant step in the model and would quickly point to a case of over-fitting. As such, the basic rules required to build the models are overlapping, with many shared elements.

To find the best-fit parameter sets for the observed NET-seq distributions for each gene, a set of 150,000 parameters were sampled uniformly via latin-hypercube sampling for the unknown parameters in each model. For each set, the processes of transcription initiation, elongation and termination were simulated for 10,000 "genes" with identical parameters for 40 minutes in increments of 0.005 minutes per time-step to allow the system to reach steady-state. The output distribution of transcriptionally-engaged polymerase was then taken as the sum of locations of polymerase across the 10,000 simulations. The best-fit parameter set was then determined as the set with lowest Kolmogorov-Smirnov statistic when compared to raw NET-seq distribution. A full description of the algorithm is given in Section 7.4.

Transcription initiation

Polymerase binding to the promoter is modelled with constant rate, i.e. the times between binding events are distributed exponentially. This assumes that polymerase is not saturated *in vivo*, or at least there is enough free polymerase in the cell to allow constant binding rates at active genes.

We add further complexity to the model by allowing promoters to exist in an active and inactive state, similar to the model presented in chapter 3. Here, a promoter is allowed to switch stochastically between an active and inactive state with constant rates; in an active state, polymerase binds with constant rate as above.

Finally, we allow the site of transcription initiation to vary between initiation events as in Azofeifa & Dowell 2017. Here, we allow the target of transcription initiation to vary relative to the 'annotated' start site. Each initiation event is given a start site sampled from a Normal distribution $N(\mu, \sigma^2)$, where μ is the target start site relative to the annotated start site, $\mu \in (-250, +250)$ and σ is the standard deviation of the start site targeting of polymerase, $\sigma \in (0, 50)$.

Promoter-proximal dynamics

To accommodate the well-established concept of promoter-proximal pausing into the models (Rougvie & Lis 1988, Zhou *et al.* 2012, Qiu & Gilmour 2017) and motivated by the 5' "peaks" of polymerase observed in NET-seq data (see Chapter 4), we incorporated a site-specific pause step into polymerase elongation. The site of proximal pausing and the rate of pause-release are sampled from uniform distributions for each iteration, allowing for long and short pauses situated proximal and distal to the TSS. Once a polymerase reaches the pause site, it will stop moving and will start moving again after an exponentially distributed amount of time, with mean sampled uniformly.

Elongation

Elongation is thought to be a reasonably smooth process under optimal conditions (Larson *et al.* 2014), where transcribing polymerase moves in approximately even steps in the absence of pausing or stalling. Here, we model the number of nucleotides processed by the polymerase per unit time as a Poisson distribution, equivalent to a constant probability of processing the next nucleotide in a small increment. This is analogous to the modelling in Chapter 3 where the time to next leaving the current nucleotide is distributed exponentially. Here, we model the probability of moving to the next nucleotide in the time interval $(t, t + \Delta t)$ as constant with probability $k\Delta t$. As such, this results in approximately constant elongation rates for each polymerase throughout the gene body, with a

variance equal to this elongation rate:

$$\#\{Nucleotides\ moved\ in\ \Delta t\} \sim Poi(k\Delta t)$$
$$\mathbb{E}[\#\{Nucleotides\ moved\ in\ \Delta t\}] = k\Delta t$$
$$Var[\#\{Nucleotides\ moved\ in\ \Delta t\}] = k\Delta t$$

Stalling and occlusion

One factor that immediately sets apart this model setup from a simple one-directional diffusion model is the occlusion of polymerases. As elongation is itself here modelled as a stochastic process, each individual polymerase moves different increments in the same amount of time, occasionally creating events where a polymerase will collide with one transcribing in front. *In vivo*, it has been proposed that under these scenarios, stalled polymerases blocking elongation of upstream polymerases are ubiquitylated and targeted for degradation (Svejstrup 2007). Once the polymerase is removed from the DNA, the upstream polymerase is free to restart elongation through the gene body.

Here we present a number of possible steps that may occur on collision of polymerases. In a model without polymerase stalling as a possible state, the polymerases will continue to progress forward, but those upstream polymerases will always be set behind the polymerase immediately downstream, through an occlusion model (Palmer *et al.* 2011). This has the potential to create 'traffic jams' of polymerases throughout the entire gene body, eventually reaching the termination point.

In a model with polymerase stalling, if upon a collision the downstream polymerase is in a moving state, the upstream polymerase is put into a stalled state while the downstream polymerase continues elongating through the gene. If, however, upon collision the downstream polymerase is stalled, the upstream polymerase will also be stalled and this is used as a signal for protein degradation, whereby in the model the downstream polymerase is removed from the gene, resulting in early termination.

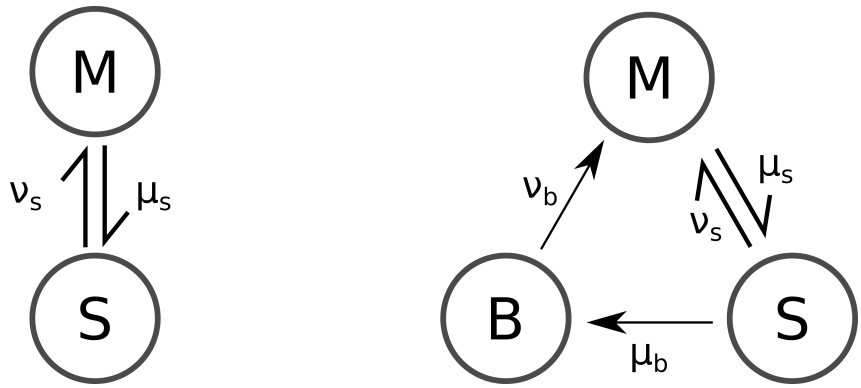


Figure 5.1: The transitions undertaken by polymerase in models with stalling (left) and models with stalling and backtracking (right). The transitions are modelled as a continuous-time Markov-chain, where the time to transition from one state to another is distributed exponentially with rates given by the arrows. States are Moving (M), Stalled (S) and Backtracked (B).

Backtracking

A known release mechanism of stalled polymerase is via backtracking and recovery of polymerases. When nucleotides are misincorporated into the RNA sequence, polymerase will pause and backtrack upstream to leave an exposed 3' end of RNA. The exposed nascent RNA is then cleaved by TFIIS (Dst1 in *S. cerevisiae*) before transcription is restarted (Izban & Luse 1992, Churchman & Weissman 2011). In the models that incorporate backtracking, we include backtracking as another kinetic state of polymerase, whereby stalled polymerase may also stochastically enter a backtracked state, from where the elongation activity can be restarted (Figure 5.1). Elongation will not start, however, if the backtracked polymerase experiences a collision with an advancing polymerase. If this occurs, the upstream polymerase will be stalled and this signals the starting point of the degradation pathway and the downstream, backtracked polymerase will undergo early termination and be removed.

Early termination

The processivity of polymerase is defined as the proportion of initiation events that result in a fully formed transcript. Estimates for the processivity of polymerases *in vivo* have a large range, with lower-bound estimates as little as 1% of initiation events reaching termination (Steurer *et al.* 2018). Early regulatory steps such as proximal-promoter pausing have been proposed as a checkpoint, whereby those elongation complexes without the full complement of capping or elongation factors would undergo early termination (Tel-

lier *et al.* 2016). Therefore to allow early termination to occur via a mechanism other than collisions of polymerase, we included termination through two additional processes. First, we allow termination to occur uniformly throughout the gene at a constant rate and secondly we model early termination via a Poisson distribution, where the mean of the Poisson distribution, taken from the closed interval $\in [1, 1000]$, gives the centre of the termination location.

Multiple kinetic regions

Finally, we wished to test the possibility that transcription through a gene may not be as simple as having the same dynamic rates at all positions along the gene. Certainly recent studies have suggested different elongation rates at the proximal and distal regions of genes (Booth *et al.* 2018, Fitz *et al.* 2018), where the DSIF component Spt5 and its post-translational modifications have been associated with controlling the speed of polymerase through different sections of the gene body.

To account for this in the model, we allowed two parameter regimes where a second set of pausing, backtracking and restarting rates were sampled independently of the first set. The size of the first window was also sampled from a uniform position, where the break-point of the first kinetic window was chosen from $[1, 1000]$. It is worth noting that a window location of 1 or 1000 is equivalent to a single kinetic window encompassing the entire "gene".

This addition to the model is novel compared to pre-existing models of polymerase activity (Choubey *et al.* 2015, Ehrensberger *et al.* 2013, Azofeifa & Dowell 2017). While the recent studies from The Lis and Pavri labs have highlighted the presence of a switch in polymerase dynamics, potentially regulated by DSIF, we are interested in whether this phenomenon is required when aiming to capture the distributions of transcribing polymerase.

5.2.2 A list of analysed models

The models used to simulate transcribing polymerase profiles fall into a number of sub-categories, the parameters and reactions found in each model are as follows:

Core Model

Polymerase initiates and elongates with occlusion

Stalling models

Core model + polymerase stochastically stalls and restarts

Core model + stalling + removal of stalled polymerases undergoing collision

Backtracking models

Core model + stalling + polymerase stochastically enters a backtracked state from a stalled state and can restart

Core model + stalling + backtracking + removal

Early termination

Core model + constant early termination

Core model + early Poissonian termination

Core model + stalling + constant early termination

Core model + stalling + early Poissonian termination

Core model + stalling + removal + constant early termination

Core model + stalling + removal + early Poissonian termination

Core model + stalling + backtracking + constant early termination

Core model + stalling + backtracking + early Poissonian termination

Core model + stalling + backtracking + removal + constant early termination

Core model + stalling + backtracking + removal + early Poissonian termination

Two windows

Core model + stalling and release rates in window 1 + stalling and release rates in window 2

Core model + early termination + stalling and release rates in window 1 + stalling and release rates in window 2

Core model + stalling, backtracking and release rates in window 1 + stalling, backtracking

and release rates in window 2

Core model + constant early termination + stalling, backtracking and release rates in window 1 + stalling, backtracking and release rates in window 2

Core model + early Poissonian termination + stalling, backtracking and release rates in window 1 + stalling, backtracking and release rates in window 2

Variable start site selection

Core model + stalling + removal + variable start site

Core model + stalling + backtracking + variable start site

Core model + stalling + backtracking + removal + variable start site

Core model + stalling in 2 windows + variable start site

Core model + stalling and backtracking in 2 windows + variable start site

Bursty promoter

All of the above models, with a promoter that stochastically switches between an active and inactive state.

5.2.3 Calculating the observed elongation rate

Calculation of the elongation rate of transcribing polymerase has remained elusive both *in vitro* and *in vivo*, with a whole host of values given from both direct and indirect observations. Certainly one may expect a discrepancy between the values calculated for polymerase on a naked DNA template *in vitro* (Larson *et al.* 2014) and through a compacted chromatin environment *in vivo* (Zenklusen *et al.* 2008, Gressel *et al.* 2017, Booth *et al.* 2018). To further complicate matters, we must take into account the multiple states of polymerase when attempting to derive an elongation rate. Indeed, we must separate the speed of polymerase from the time a polymerase takes to proceed from one end of the gene to the other, which is arguably of more use to us.

A better metric of polymerase velocity may be an "observed" elongation rate, where the time polymerase is static in a stalled or backtracked state is taken into account. The observed elongation rates will vary for the models analysed in this chapter, as different

models have different numbers of dynamic states and as such we require at least three measures of "observed" elongation rate.

In a model without stalling or backtracking, we can simply output the elongation rate of polymerase, given by k above. By taking standard results from Markov theory, we can calculated the expected "observed" elongation rates by considering the states exhibited by a polymerase under the various models (Figure 5.1). In a model with stalling, a polymerase transitions from a moving to a stalled state and back again with rates μ_s and ν_s, respectively. The time a polymerase stays in a moving state is distributed exponentially with parameter μ_s. As such the expected time to leave each state is given by:

$$\begin{aligned}\mathbb{E}\left[T_M\right] &= \frac{1}{\mu_s} \\ \mathbb{E}\left[T_S\right] &= \frac{1}{\nu_s}\end{aligned}$$

and the fraction of time the polymerase is moving is given by:

$$\begin{aligned}F_M &= \frac{\frac{1}{\mu_s}}{\frac{1}{\mu_s} + \frac{1}{\nu_s}} \\ &= \frac{\nu_s}{\mu_s + \nu_s}\end{aligned}$$

where k is the actual elongation rate of the polymerase in kb/min. Therefore the observed elongation rate for this model is given by:

$$k_{obs} = \frac{k\nu_s}{\mu_s + \nu_s} \tag{5.1}$$

Under a model with backtracking, the residence times in each state are slightly different. The expected times to leave each state are given by:

$$\begin{aligned}\mathbb{E}\left[T_M\right] &= \frac{1}{\mu_s} \\ \mathbb{E}\left[T_S\right] &= \frac{1}{\nu_s + \mu_b} \\ \mathbb{E}\left[T_B\right] &= \frac{1}{\nu_b}\end{aligned}$$

To consider the fraction of time spent in a moving state, we must consider the time to leave the moving state immediately after arrival and the mean return time to the moving state immediately after leaving. Conditioning on the steps after leaving the moving state gives us the following:

$$\begin{aligned}
\mathbb{E}\left[T_{S\to M}\right] &= \mathbb{E}\left[T_{S\to M}|S\to M\right]\mathbb{P}\left\{S\to M\right\} + \mathbb{E}\left[T_{S\to M}|S\to B\right]\mathbb{P}\left\{S\to B\right\} \\
&= \frac{1}{\nu_s+\mu_b}\frac{\nu_s}{\nu_s+\mu_b} + \frac{\mu_b}{\nu_s+\mu_b}\left(\frac{1}{\nu_s+\mu_b} + \mathbb{E}\left[T_{B\to M}\right]\right) \\
&= \frac{1}{\nu_s+\mu_b}\left(\frac{\nu_s+\mu_b}{\nu_s+\mu_b} + \frac{\mu_b}{\nu_b}\right) \\
&= \frac{1}{\nu_s+\mu_b}\left(1+\frac{\mu_b}{\nu_b}\right)
\end{aligned}$$

As the mean time to leave the moving state is given by $\frac{1}{\mu_s}$ as with the previous model, the fraction of time spent in the moving state is given by:

$$\begin{aligned}
F_M^* &= \frac{\frac{1}{\mu_s}}{\frac{1}{\nu_s+\mu_b}\left(1+\frac{\mu_b}{\nu_b}\right)+\frac{1}{\mu_s}} \\
&= \frac{\nu_s+\mu_b}{\mu_s\left(1+\frac{\mu_b}{\nu_b}\right)+\nu_s+\mu_b}
\end{aligned}$$

and the observed elongation rates is given by:

$$k_{obs}^* = k\frac{\nu_s+\mu_b}{\mu_s\left(1+\frac{\mu_b}{\nu_b}\right)+\nu_s+\mu_b} \tag{5.2}$$

The same result is also derived by considering the stationary distribution, which is found by calculating the stationary vector ξ, given by:

$$\xi Q = 0$$

and the Q-matrix is given by the rates between each state as in Figure 5.1:

$$Q = \begin{pmatrix} -\mu_s & \mu_s & 0 \\ \nu_s & -(\nu_s+\mu_b) & \mu_b \\ \nu_b & 0 & -\nu_b \end{pmatrix}$$

The stationary distribution is given by the row-vector which satisifes:

$$\xi Q = 0 \tag{5.3}$$

$$\Sigma_i \xi_i = 1 \tag{5.4}$$

$$-\mu_s \xi_1 + \nu_s \xi_2 + \nu_b \xi_3 = 0 \tag{5.5}$$

$$5.3 \implies \mu_s \xi_1 - (\nu_s + \mu_b)\, \xi_2 = 0 \tag{5.6}$$

$$\mu_b \xi_2 - \nu_b \xi_3 = 0 \tag{5.7}$$

Taking equations 5.7 and 5.5 gives the following:

$$\xi_2 = \frac{\nu_b}{\mu_b} \xi_3 \tag{5.8}$$

$$\xi 1 = \frac{\nu_b}{\mu_s} \left(\frac{\nu_s}{\mu_b} + 1 \right) \xi_3 \tag{5.9}$$

Combining these gives:

$$\xi = \left[\frac{\nu_b}{\mu_s} \left(\frac{\nu_s}{\mu_b} + 1 \right) \xi_3, \frac{\nu_b}{\mu_b} \xi_s, \xi_3 \right] \tag{5.10}$$

Then:

$$5.4 \implies \xi_3 \left(1 + \nu_b \left(\frac{\nu_s}{\mu_s \mu_b} + \frac{1}{\mu_s} \right) + \frac{1}{\mu_b} \right) = 1 \tag{5.11}$$

$$\implies \xi_3 = \frac{\mu_s \mu_b}{\mu_s \mu_b + \nu_b \left(\nu_s + \mu_b + \mu_s \right)} \tag{5.12}$$

Therefore the long-run proportion of time the state spends in state 1 (moving) is given by:

$$\xi_1 = \frac{\nu_b}{\mu_s} \left(\frac{\nu_s}{\mu_b} + 1 \right) \frac{\mu_s \mu_b}{\mu_s \mu_b + \nu_b \left(\nu_s + \mu_b + \mu_s \right)} \tag{5.13}$$

$$= \frac{\nu_b \left(\nu_s + \mu_b \right)}{\mu_s \mu_b + \nu_b \left(\nu_s + \mu_b + \mu_s \right)} \tag{5.14}$$

$$= \frac{\nu_s + \mu_b}{\mu_s \left(1 + \frac{\mu_b}{\nu_b} \right) + \nu_s + \mu_b} \tag{5.15}$$

5.3 Fitting simulated distributions to NET-seq

The aims of fitting to raw data were two-fold: firstly we wished to establish which model determines the best fit of transcription dynamics and secondly we wished to see how the dynamics of transcription in the best-fit model changed across the genome and upon intervention.

For each model and for each gene, we determined the best fit as the set of parameters that gave the minimal Kolmogorov-Smirnov statistic (Massey 1951). The likelihood for the data under this set of parameters was taken to be the maximum likelihood statistic for the given gene under the current model. Repeating this process for every gene enabled us to calculate the Akaike Information Criteria:

$$AIC = 2k - 2ln\left(\hat{L}\right)$$

where k is the number of unknown parameters in each model, n is the size of the dataset, here the number of genes analysed and $\hat{L}$ is the maximum likelihood for the dataset under the given model.

The AIC is has been found to over-fit in the case of small sample size. Therefore Hurvich & Tsai 1989 proposed the corrected Akaike Information Criteria, with a correction term for small sample sizes:

$$AICc = 2k - 2ln\left(\hat{L}\right) + \frac{2k^2 + 2k}{n - k - 1}$$

In the datasets analysed in this chapter, we have ~ 1000 observations, given by the signal for each gene, and $k < 15$ unknown parameters. The correction term is thus on the order of 0.1:

$$\begin{aligned}\frac{2k^2 + 2k}{n - k - 1} &\sim \frac{O\left(10^2\right)}{O\left(10^3\right)} \\ &\sim O\left(10^{-1}\right)\end{aligned}$$

As differences in AIC values between two models in the range of $0-2$ are not considered enough evidence to reject one model over another (Burnham & Anderson 2004), we adopt

the use of the simpler AIC for the remainder of the chapter.

It is assumed that there is a degree of structural un-identifiability in the model. Indeed, results from chapter 3 indicate that the total number of polymerases on a gene at one time is defined as the ratio of initiation rate to elongation rate, therefore we are not able to determine a single set of parameters that best explain transcription dynamics at each gene. We took the 100 best parameter sets ($< 0.1\%$), from the 150,000 sampled, as determined by the Kolmogorov-Smirnov statistic to be the sets of parameter that defined the dynamics at each gene. Each group of parameter sets was then used to determine the initiation-elongation ratio, mean moving time and other statistics, such as early termination rate, for each gene.

The data itself exhibits a large degree of noise, thus while we are able to obtain maximum likelihood estimates for each gene, in order to gain a clearer picture of the ranges of dynamics we turned to the sets of parameters for sub-clusters of the genome. By grouping data into clusters like this, we are not able to gain a full picture of the dynamics of transcription at each gene, but instead gain a much higher-level view of how representative profiles across the genome (See Figure 4.3b) are related to the underlying dynamics. In particular it gives us an idea of how much heterogeneity there exists in transcriptional dynamics across the genome of *S. cerevisiae*.

Our results have a degree of similarity to those found by modelling PolII ChIP data from *Arabidopsis thaliana* in Wu *et al.* 2016. The simulated profiles of transcribing polymerases exhibit large spikes in signal immediately downstream of the TSS. This signal is likely to be made up of early transcribing RNA PolII, with very short nascent RNA in the active site. As such, these short nascent strands would not be taken forward for sequencing, as the number of nucleotides would be too short to uniquely align to the genome (< 15 nt in *S. cerevisiae*, < 30 nt in HeLa cells). These "spikes" of polymerase signal have been kept in the simulation figures as they do highlight the effects of changes in dynamics in Figures 5.8, 5.10 & 5.12 on the polymerase build-up in the promoter-proximal regions.

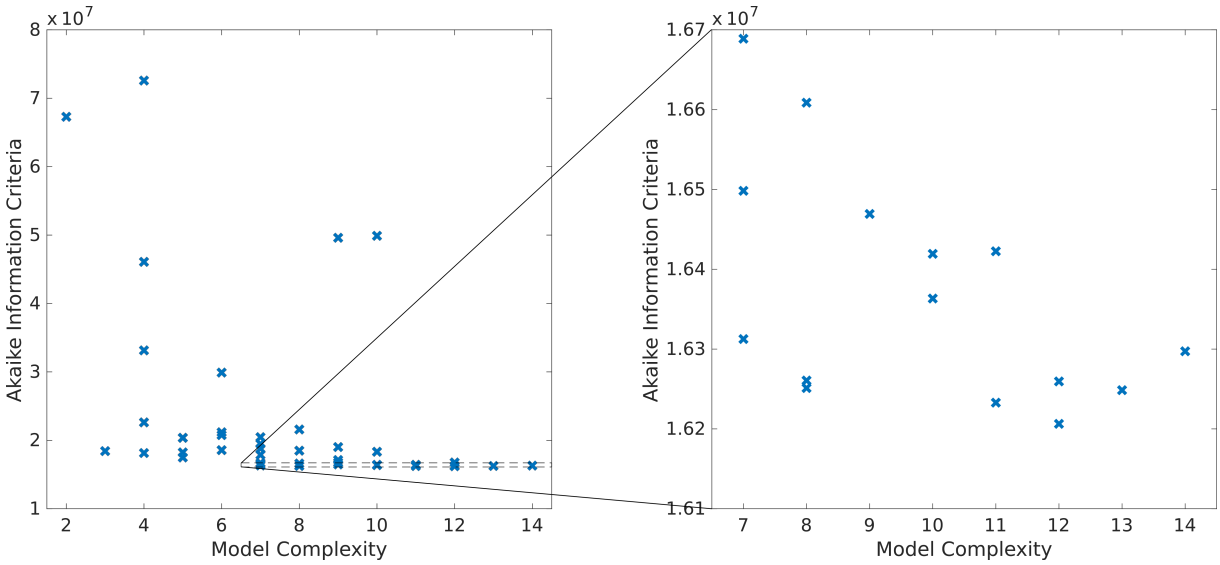


Figure 5.2: Akaike Information criteria (AIC) for the tested models after fitting to *S. cerevisiae* wild-type NET-seq data from Fischl *et al.* 2017. Each cross represents a tested model and models are ordered by number of free parameters (model complexity) with an enlargement of the marked area shown on the right. The best-fit model with minimal AIC is shown to exist with 12 free parameters.

5.4 Backtracking and two windows are components of the best-fit model

We determined the polymerase distributions by fitting the synthetic polymerase distributions to a wild-type *S. cerevisiae* NET-seq experiment from Fischl *et al.* 2017 (See Figure 4.3 in Chapter 4). Assessing fit to the data by the AIC, we determined that the best-fit model (Figure 5.2) to take forward was a model including:

- Stochastic stalling
- Stochastic backtracking
- Removal of polymerase following collisions with backtracked polymerase
- Two kinetic windows, with different stalling and backtracking rates
- Early Poissonian termination

Using this approach of maximum likelihood estimation followed by calculation of information criteria, we were able to determine that models incorporating bursting promoters, fixed-site polymerase pauses, constant rates of early termination and variable start-site

selection resulted in poor or over-fitted distributions. This does not rule out the possibility that these reactions occur *in vivo* under normal transcription. Indeed, in chapter 3 we argued that a bursty model of transcription is required to explain the distributions of nuclear and cytoplasmic RNA seen in *S. cerevisiae* as has been shown previously (Zenklusen *et al.* 2008, Lenstra *et al.* 2015, Ietswaart *et al.* 2017). Here, we would argue that NET-seq is not a good technique to study promoter bursting, given the measurements taken are an average of distributions across a population of cells, we might expect any burst-like patterns to be smoothed-out through averaging the distributions of polymerase across a population of >1,000,000 cells.

The model identified here as best capturing the behaviour of transcribing polymerase in *S. cerevisiae* also identified stalling as a constant process occuring throughout the gene body, rather than occuring at a fixed-point. This does not rule out the existence of the well-documented promoter-proximal pause, but here we would argue that this pause may in fact exist as a region of high pausing near the promoter, as opposed to pausing at a single nucleotide. This points to polymerase elongation being a much less smooth process, where a modelling framework requires the multi-state polymerase incorporating the coupled stalling and backtracking reactions.

That polymerase backtracking forms a core part of the required model is in accordance with similar *in silico* analysis of theoretical transcription carried out in Li & Zhang 2015. In that study, they argue that to overcome any blockages to transcription, polymerase must either undergo backtracking, termination or bypass blockages altogether in order to maintain high efficiency of transcription. While that study only took transcriptional efficiency as a desired output, here we are able to couple this to *in vivo* measurements of polymerase activity and relate this to the models analysed in Li & Zhang 2015 and Choubey *et al.* 2015.

Finally, it is perhaps not surprising that models are better fit with early termination occurring around a fixed point than constantly throughout the gene. The polymerase has been shown to be slowed while passing around nucleosomes (Churchman & Weissman 2011) and if obstacles are on the DNA, they can interfere with the the polymerase and even cause transcription not to proceed beyond this point (Lenstra *et al.* 2015). As

many factors targeting the nucleosomes or specific DNA sequences bind the DNA, the polymerase will have to negotiate these obstacles and in extreme circumstances undergo abortion. A possible further model to test in the future may be early termination occurring at more than one fixed position across the gene. This would either confirm the existence of only two dynamic windows partitioning the genes into promoter-proximal and -distal, or introduce a higher number of dynamic shifts which the polymerase undergoes throughout elongation.

5.4.1 Relating parameters to polymerase distributions

To begin to unpick the complexity of the distributions of polymerase and the relations to the dynamics, we sought to determine which parameters are the best determinants for key statistics when analysing nascent transcription datasets (Fischl *et al.* 2017, Churchman & Weissman 2011, Nojima *et al.* 2015). For the model with lowest AIC when fitting to *S. cerevisiae* NET-seq data, we took the values of initiation/elongation ratio, first window mean moving time, second window mean moving time, window location and early termination location and correlated them with total polymerase levels (Figure 5.3) and proximal-distal ratio (Figure 5.5).

The total amount of polymerase seen via a technique such as NET-seq or PolII ChIP-seq is strongly determined by the ratio of initiation rate to elongation rate as determined by the simulations of transcription (see Figure 5.3, Ehrensberger *et al.* 2013). This result does not come as a surprise, taking into account results from queueing theory, where here we expect elongation of a gene to occur on a longer time-scale than initiation, resulting in a queue of polymerase along a gene. We found, however, that the mean moving time of polymerases is not as strong a determinant of PolII levels, with a moderate correlation suggesting that more time moving in the first window corresponds to more steady-state polymerase on the gene and vice-versa in the second window. We further saw no correlation between the size of the first window and total polymerase on the gene or the location of any early termination events.

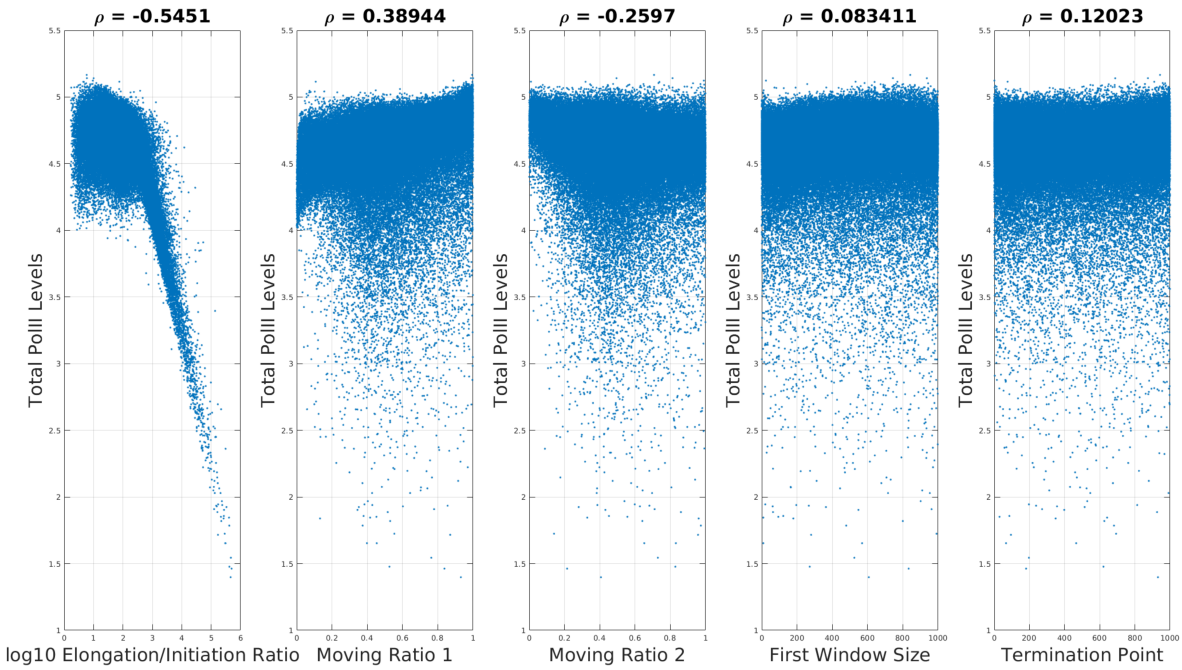


Figure 5.3: Correlations between parameter values and total number of polymerase molecules on a gene at steady-state. Values shown are taken from synthetic distributions of polymerase transcription, correlation coefficient is Spearman's ρ.

NET-seq simulation observations are consistent with modelling RNA FISH data

An important check to make before we proceed with additional analysis is whether we are introducing any biases into our inference by using different experimental techniques and different models to infer dynamics about transcription. We are not able to directly compare the modelling outputs from the simpler model of RNA production in Chapter 3 with the more high-resolution polymerase elongation model presented here. However, we have determined that the total amount of polymerase on a gene at any one time is a factor of both the initiation rate and the elongation rate of polymerase (Figure 5.3), both of which we were able to infer from the model of RNA FISH data. Therefore to determine the validity of our two approaches to study the dynamics of transcription initiation and elongation, we compared those dynamics inferred from the FISH model with the raw NET-seq data studied in this chapter.

We found a strong negative correlation between the elongation/initiation ratio for the genes studied via RNA FISH and the number of NET-seq reads at each gene. Here, a shorter, fixed region was used to account for the different lengths of the genes and to nullify any effects which would be brought about by the shapes of the distribution of

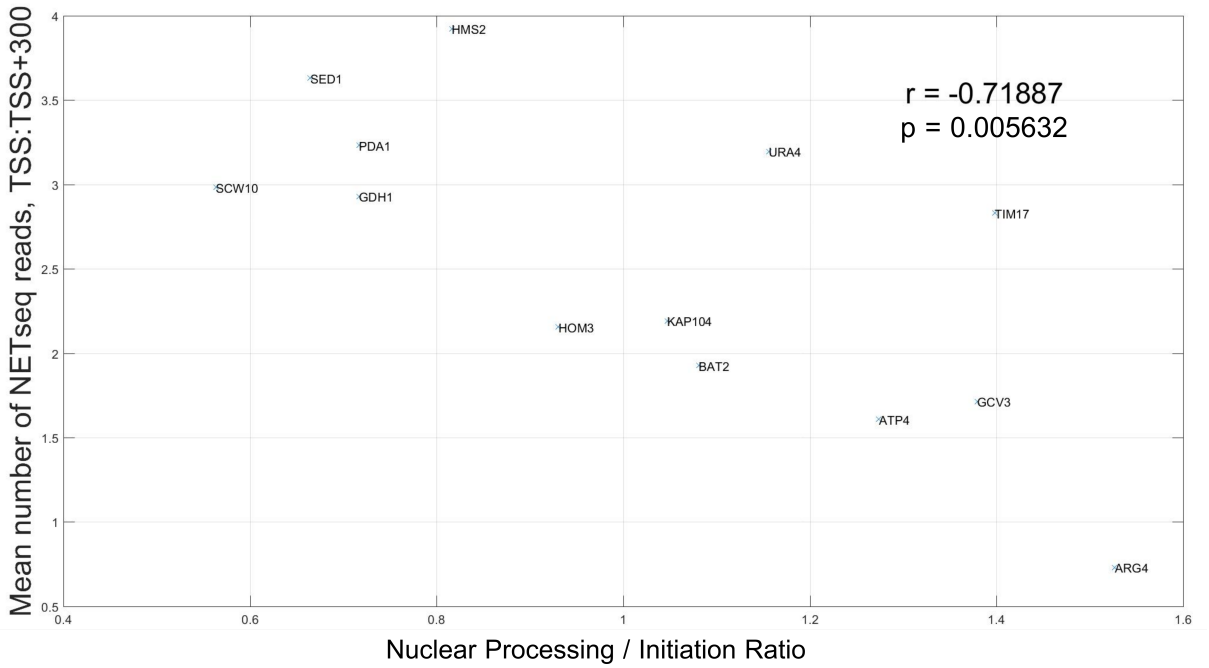


Figure 5.4: The mean number of NET-seq reads at each of the 13 analysed genes in glucose media compared with the ratio of the inferred mean production rate and nuclear processing rate. Shown is the Pearson correlation coefficient.

polymerase at each gene.

This strongly indicates that the rates deduced by modelling RNA FISH data are comparable with those deduced by modelling NET-seq data and that the total amount of nascent transcription occurring is a function of many dynamic processes at each gene. Therefore we believe that an experimental approach whereby the amount of global antisense transcription is altered may not have an immediately obvious effect on the NET-seq distributions as the initiation and elongation rates were both found to change in the same manner in Chapter 3, suggesting that this initiation/elongation ratio would remain the same. Similarly, we may observe the same effect by affecting the modifications to the chromatin as we saw the same effect after inhibition of the histone de-acetylase Set3.

Transcription parameters dictate the distribution of RNA PolII across the gene

As stated above, the levels of polymerase at each gene give us an indication of the transcription occurring, but we can glean further information by looking at the distributions of polymerase. As a starting point, we turned to the ratio of polymerase found in the 5' and 3' ends of genes, termed the proximal and distal regions of the gene, respectively. This

metric has taken on multiple names as the pause ratio (Tastemel *et al.* 2017, Booth *et al.* 2018), travelling ratio (Rahl *et al.* 2010) or processivity ratio (Malik *et al.* 2017). Here, we will show that while this ratio is related to the degree of pausing in the promoter-proximal region of the gene, it is perhaps a better indication of the processivity of a gene, again itself a function of pausing.

The identification of promoter-proximal pausing in metazoa drove forward the idea that an accumulation of polymerase in the proximal regions of genes could indicate more 5' pausing of polymerase (e.g. Booth *et al.* 2016, Chen *et al.* 2015). Here, we wished to test which parameters resulted in distributions with greater proximal-biases of polymerase.

We identified the initiation/elongation ratio and the mean first window moving time as key determinants of the ratio of proximal-to-distal polymerase at steady-state (Figure 5.5). Perhaps rather than being indicative of higher pausing, a build-up of proximal polymerase is suggestive of polymerases unable to clear the promoter region, either because the rate of initiation is too high, or because there are obstacles the polymerase must pass in this region.

Finally, we investigated the role of transcription dynamics on the processivity of polymerase. We defined transcription processivity as the percentage of "successful" transcription initiation events, i.e. events in which the polymerase reaches the end of the gene, or beyond 1,000bp after the TSS here (see Figure 5.6). We found that the initiation-elongation ratio is again a good determinant of the processivity of transcription, with relatively fast elongation resulting in more efficient transcription. Furthermore, we found that the mean second window moving time was also a reasonably strong determinant of transcript processivity, reflecting the lack of any distal barriers to elongation resulting in more efficient transcription events.

We will use the NET-seq profiles obtained from mutants in *S. cerevisiae* and drug-treated HeLa cells shown in Chapter 4 to investigate how interfering with the transcriptional network changes model fits. This will inform how individual factors contribute to the dynamics of the polymerase reading the DNA template and begin to bring together how all of these factors either work together or antagonistically.

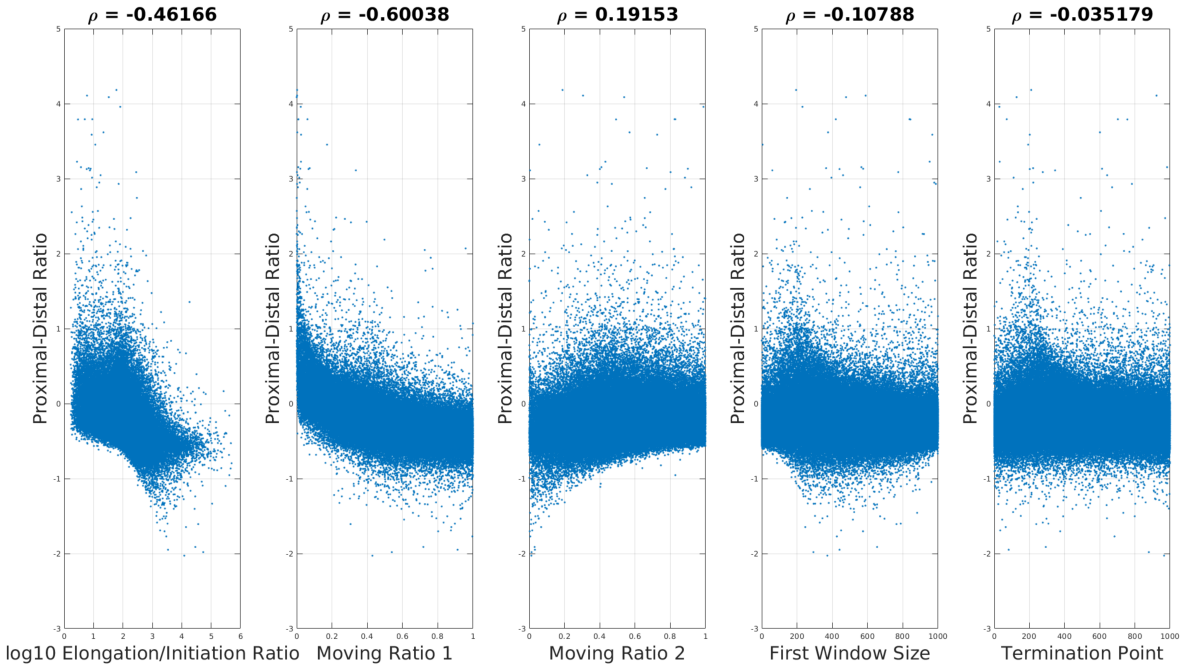


Figure 5.5: Correlations between parameter values and the proximal-distal ratio of polymerase on each gene. The proximal-distal ratio is defined as the total number of molecules in the first 200bp after the TSS divided by the total number of reads 201–1,000bp after the TSS. Values shown are taken from synthetic distributions of polymerase transcription, correlation coefficient is Spearman's ρ.

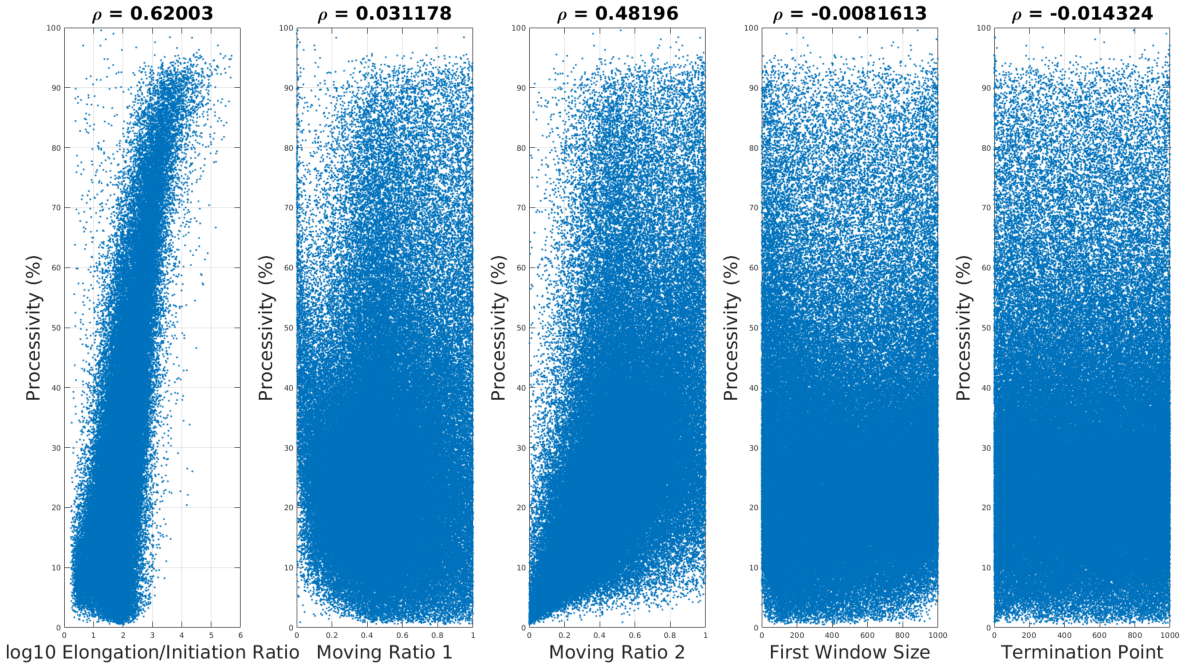


Figure 5.6: Correlations between parameter values and the processivity of polymerase at each gene. Processivity is defined as the the percentage of initiation events in which the polymerase reaches the end of the gene (here, 1,000bp downstream of the TSS). Values shown are taken from synthetic distributions of polymerase transcription, correlation coefficient is Spearman's ρ.

5.4.2 Sub-groups of genes are best-fit by different models of transcription

We wished to determine whether the best-fit model was consistent regardless of the sub-groups of genes taken forward into the analysis. To do this, we took those genes clustered by k-means in Chapter 4 (Figure 4.3b) and repeated the procedure of finding the maximum likelihood and calculating the AIC. We found that the three clustered groups were best-fit by different models and one group had the same best-fit model as the entire population as determined above. The cluster which most closely follows the metagene profile of a broad peak at the 5' end of the gene (cluster 1) was best-fit by a model with a non-bursty promoter, two dynamic windows, stochastic stalling, early termination around a mean location and no removal of polymerases upon collisions. The cluster with the sharpest peak of 5' polymerase (cluster 2) was best-fit by a model with a non-bursty promoter, stochastic stalling, early termination around a mean location and with removal upon collision. The final cluster, which exhibited a flatter profile of polymerase (cluster 3) was best-fit by the same model as the entire genome (See Figure 5.7 for AIC plots for each cluster).

That the structures of the models found to fit these sub-groups of genes differ may not be surprising. It has been shown that elongation factors preferentially bind to some genes over others to influence the downstream pathways of the corresponding RNA, but this is also likely to affect the progression of the polymerase along the genes (Fischl *et al.* 2017, Battaglia *et al.* 2017). However, two of these models are not that distinct from each other as the model from cluster 2 can be thought of as a sub-group of the model from fitting to the entire genome with the backtrack rate set to 0. In this model, the polymerase still exists in moving and static states, yet the release dynamics are altered, with the polymerase undergoing a one-step, rather than a two-step, transition back to a moving state after beginning stalling.

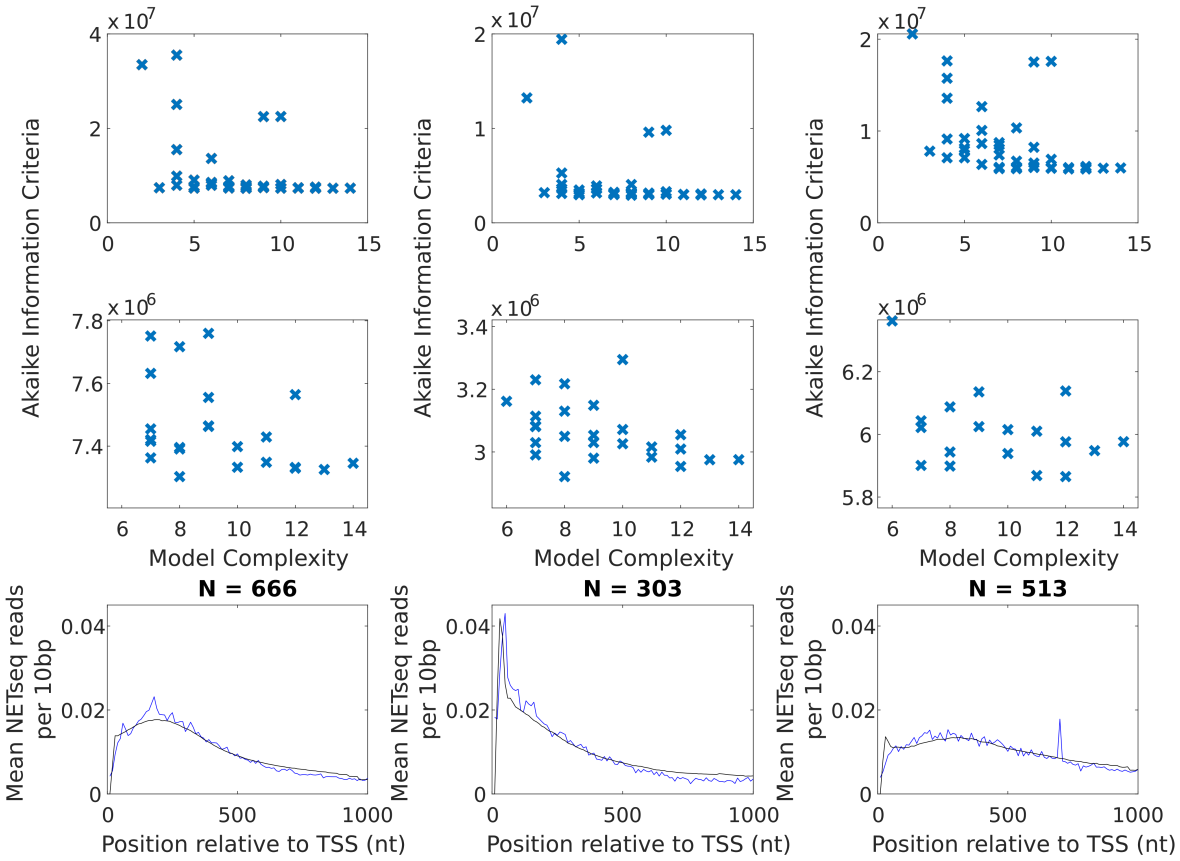
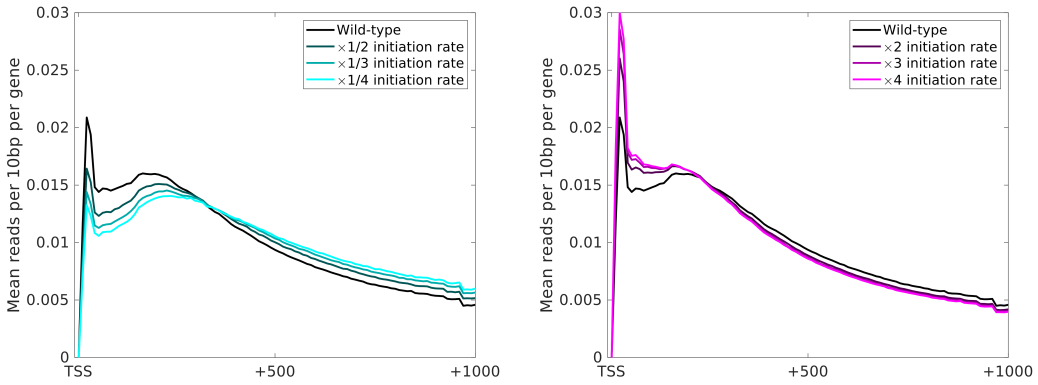


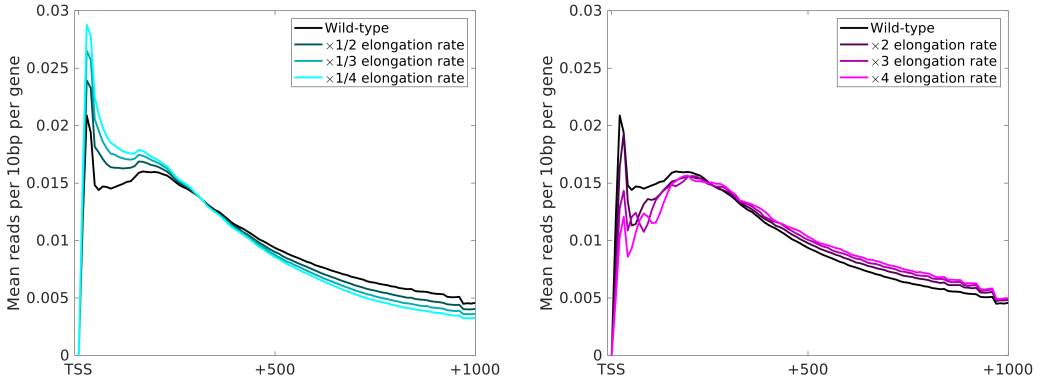
Figure 5.7: Akaike Information Criteria (AIC) for the tested models after fitting to the clusters of genes as given in Figure 4.3b. Shown are the AIC values for each cluster (top) and an enlarged area showing those models with lowest AIC values for each cluster. The models with lowest AIC values have 8 free parameters for the first 2 clusters and 12 for the last cluster, this model is the same as that determined in Figure 5.2.

5.5 Transcription dynamics dictate polymerase distributions

Using this approach of tracking the locations of transcribing polymerase and modelling the dynamics of transcription we ultimately wish to be able to determine the dynamic changes that create these distributions and how individual components of the transcription machinery contribute to these dynamics. First we wish to be able to describe any changes in polymerase distributions in terms of increases or decreases in simple transcription kinetics. To determine the effect of initiation on polymerase distributions we found the sets of parameters which produced the best-fit to the raw wild-type NET-seq data taken from Fischl *et al.* 2017 and repeated the simulations with the same parameter values and an altered initiation rate. The initiation rate was varied from 1/4 of the fitted value up



(a)



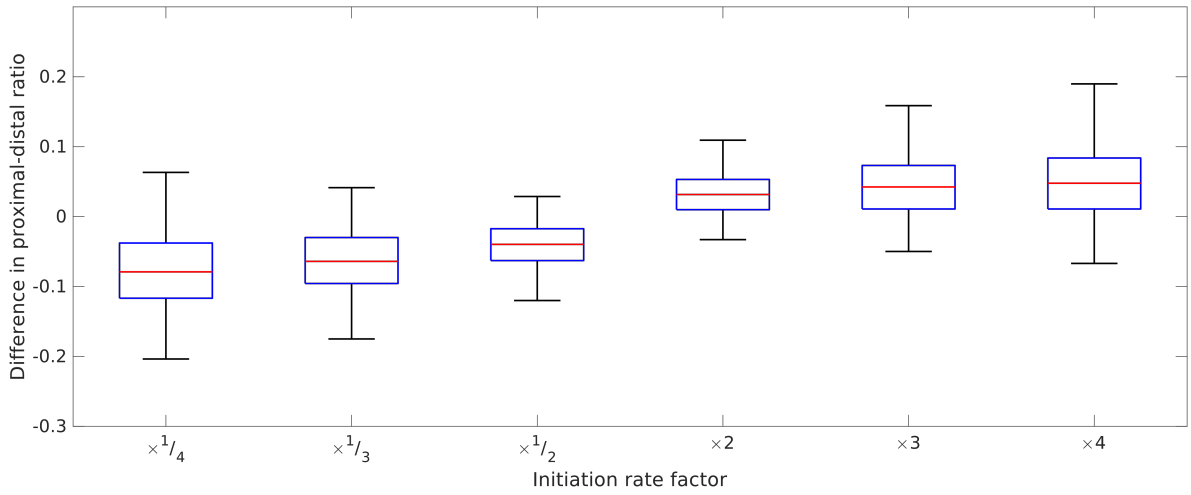
(b)

Figure 5.8: Metagene profiles of simulated NET-seq distributions after fitting to wild-type (black lines in 5.8a & 5.8b). Shown are the metagenes profiles for simulated profiles after those dynamics which fit the wild-type profiles from Figure 4.3 were simulated with either altered initiation rate (5.8a) or elongation rate (5.8b).

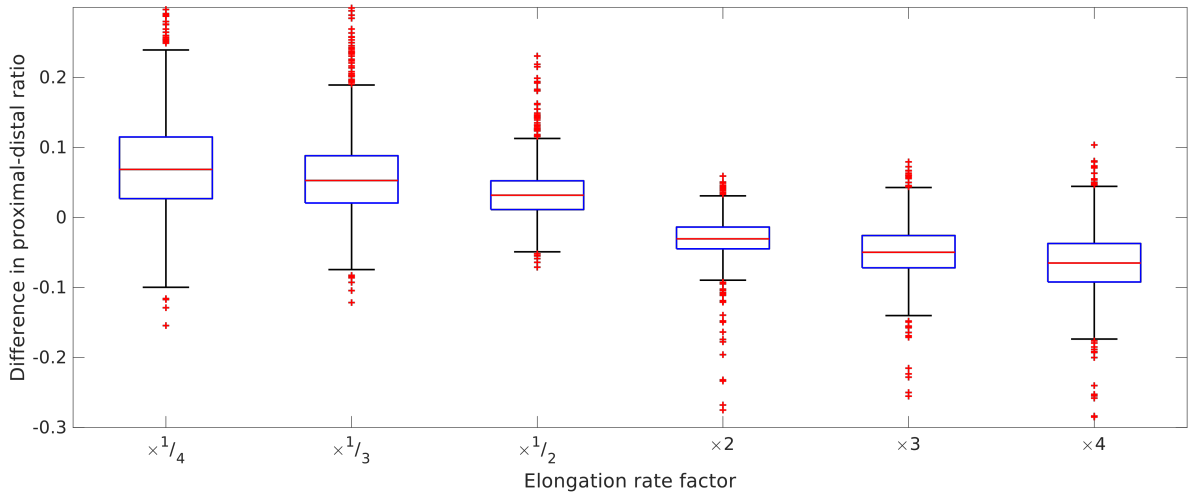
to 4× the determined value for each gene (Figure 5.8a).

The resulting distributions show that the initiation rate contributes to the ratio of polymerase found proximal to the promoter relative to the gene body (Fig 5.9). However, repeating the simulations altering the elongation rate showed that at the metagene level, the changes of initiation rate were indistinguishable from changes in elongation (Figure 5.8b). Indeed, increasing the initiation rate of transcription mirrored the profiles produced by decreasing the elongation rate by roughly the same factor. Therefore, similar to those conclusions drawn from Figure 5.3 and in Ehrensberger *et al.* 2013, it is not the absolute values of these parameters that have the greatest effect on polymerase distributions, but the ratio between initiation rate and elongation rate.

As the observed elongation rate of the polymerase is not only affected by the elongation rate of the polymerase complex, but also the rate of stalling/pausing and backtracking



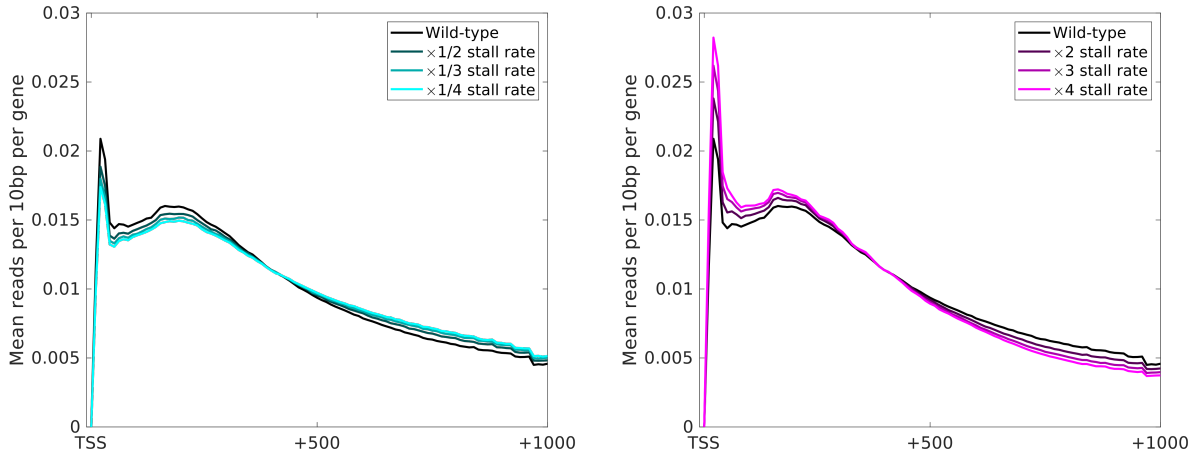
(a)



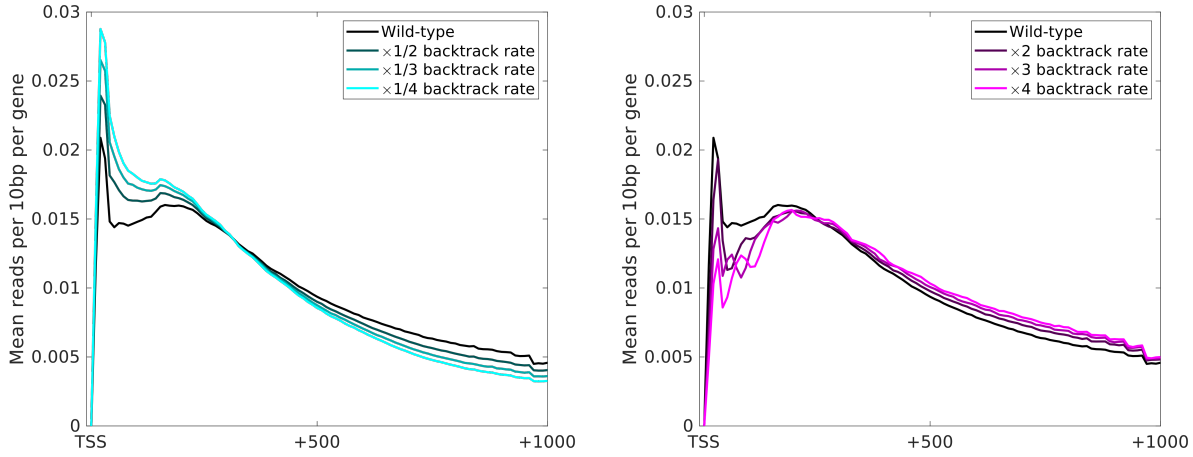
(b)

Figure 5.9: The difference in the proportion of polymerases in the first 250bp relative to the following 750bp. Shown are boxplots of the difference at each gene between the wild-type proximal-distal ratio and the simulated distributions after altering the initiation rate (5.9a) and the elongation rate (5.9b) as in Fig 5.8).

throughout the gene (Larson *et al.* 2014, Churchman & Weissman 2011, Saponaro *et al.* 2014 and Equation 5.2), we wished to test the effect of altered stalling and backtracking rates on the polymerase distributions. As above, we took the sets of parameters found to best-fit the wild-type NET-seq distributions from Fischl *et al.* 2017 and re-ran the simulations for the same parameters, but with either the stalling rates or backtracking rates for the two windows altered by factors or 2, 3 or 4. We observed that altering the stalling rate did not have a profound effect on the genome-wide averaged profiles (see Figure 5.10a) and any small effects mirrored those observed by slowing the elongation rates as shown in Figure 5.8. The general trend shows that increased stalling results in a



(a)



(b)

Figure 5.10: Metagene profiles of simulated NET-seq distributions after fitting to wild-type (black lines in 5.10a & 5.10b). Shown are the metagenes profiles for simulated profiles after those dynamics which fit the wild-type profiles from Figure 4.3 were simulated with either altered stalling rate (5.10a) or backtracking rate (5.10b).

slight shift towards a higher proportion of polymerase in the proximal region of the gene compared to the body and vice-versa.

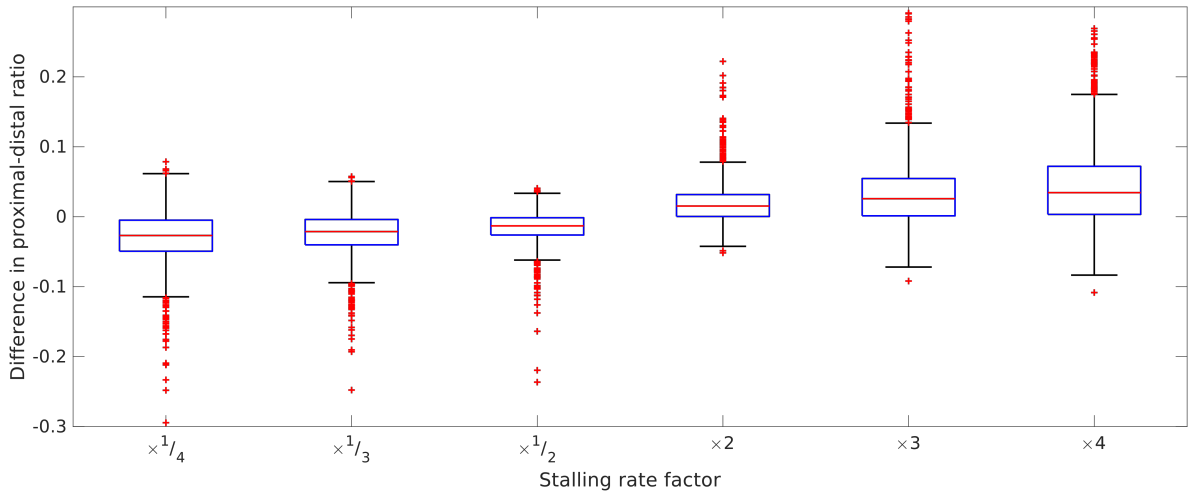
In contrast, varying the backtracking rates of the polymerase altered the distributions of transcribing polymerase to a greater degree (Figure 5.10b). Varying the backtracking rates by a factor of 2 results in large redistributions of the polymerase, with higher backtracking rates resulting in less polymerase accumulating in the promoter-proximal region of the gene and vice-versa. While stalling and backtracking are known steps the polymerase undergoes, it appears the effects of altering stalling may be more subtle than simply slowing the polymerase. We would speculate that the rate at which the polymerase stalls is not the most important determinant of the outcome of polymerase activity, but

the release rate of these stall events. Backtracking serves as a path of release for the polymerase from stalled events, and if the polymerase quickly transfers to a backtracked state before being released, this will reduce the time the polymerase spends in a static state. This would certainly explain why decreasing the rate of backtracking results in such a large accumulation of polymerase in the promoter-proximal region (Figure 5.10b). The three-state polymerase studied here presents an opportunity for different factors to act upon the polymerase while in different states and have different effects on the dynamics of the polymerase. The backtracked state of polymerase may in fact be very transient, with factors such as TFIIS (Dst1) acting rapidly to alleviate the polymerase from this state, whereas release of stalled or paused events may occur over a much longer time-scale, with much larger impact on the elongation activity of the polymerase.

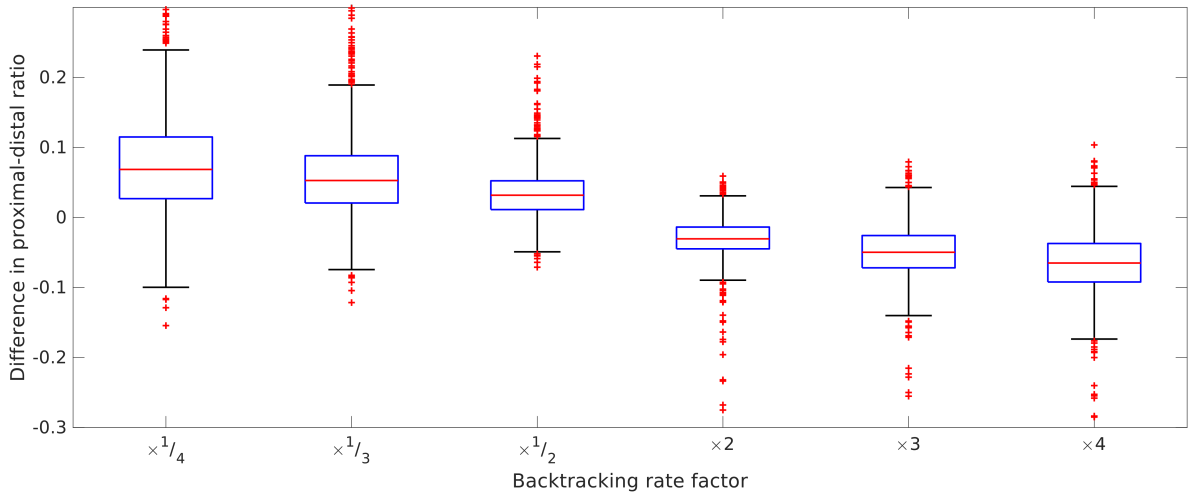
Finally, we wished to investigate the effects of extreme polymerase dynamics on the profiles of polymerase. Shown in Figure 5.12 are the polymerase profiles after the backtrack release rates and backtracking rates were set to zero (respectively, an infinite backtrack release time and a two-state polymerase model). We found that without release of backtracks, the polymerase accumulates almost entirely in the promoter-proximal region of the gene, with almost no polymerase evident at the 3' end of the gene. Without a backtrack step, we observed a moderate increase in the proportion of promoter-proximal polymerase, perhaps reflecting a dynamic regime where the polymerase spends proportionally more time in a stalled state, reflecting dynamics shown in Figure 5.10a. As we are able to interrupt backtrack release of polymerase through genetic intervention, we wish to see whether the observed polymerase *in vivo* reproduce the *in silico* profiles obtained here.

5.5.1 Altered distributions reflect altered processivity

As above, we can calculate processivity by counting the proportion of polymerases that reach the end of the gene rather than undergo early termination during the simulations. For the sets of simulated profiles above (Figs 5.8, 5.10 & 5.12) we observe that the processivity of a gene closely follows the proximal-distal ratio observed of polymerase across genes (Figure 5.13). By increasing the initiation rate or decreasing the polymerase elonga-



(a)



(b)

Figure 5.11: The difference in the proportion of polymerases in the first 250bp relative to the following 750bp for simulated NET-seq distributions. Shown are boxplots of the difference at each gene between the wild-type proximal-distal ratio and the simulated distributions after altering the stalling rate (5.11a) and the backtracking rate (5.11b) as in Fig 5.10).

tion rate we observed a more pronounced 5' peak (Figure 5.8) and a higher proximal-distal ratio (Figure 5.9). Furthermore, in simulations using these parameter sets, we observe a lower processivity, with the majority of genes exhibiting more early termination events (Figure 5.13) when initiation rates were increased or elongation rates were decreased. Correspondingly, we observed the opposite result in the simulations with slower initiation or faster elongation. In these scenarios we found genes to be more processive.

We observed a similar phenomenon when altering the stalling rate of the polymerase, with increased stalling rates resulting in decreased processivity. However, we did not observe a change upon altering backtracking rates, even though we observed a large

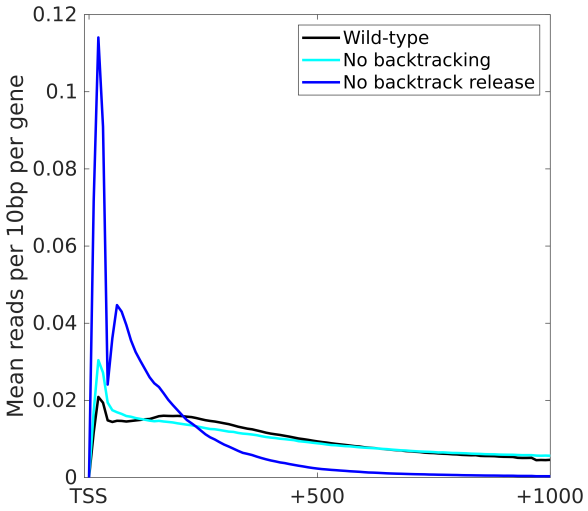


Figure 5.12: Metagene profiles of simulated NET-seq distributions after backtracking parameters were altered. Shown are the simulated profiles for parameters fit to wild-type polymerase distributions in black, the distributions after backtrack release rates were set to zero in blue and distributions with backtrack rates set to zero in cyan.

difference in the averaged profiles of polymerase. This result was further confirmed by taking the extreme results of having no backtracking at all, which did not have a consistent effect on the processivity of genes. However, we found that when backtrack release was set to zero, there was a dramatic drop in the processivity of genes, with a large proportion dropping to 0% processivity. This further demonstrates that backtracking itself may not be a key determinant of transcriptional output, as the polymerase will still enter stalled states at the same rate once backtracking events have been recovered, it is rather the rate at which backtracking or stalled events are recovered, determining the mean time the polymerase spends in a static rather than a moving state.

5.6 Transcription dynamics are distinct across the genome

To determine the range of dynamic transcription parameters that exist across the *S. cerevisiae* genome, we took the three clusters determined above for *S. cerevisiae* NET-seq data (Figure 4.3b) and separated the inferred parameters into these groups. The three clusters qualitatively split the genes into (i) a cluster which matches the genome-wide average profile of a broad peak of polymerase immediately downstream of the TSS which

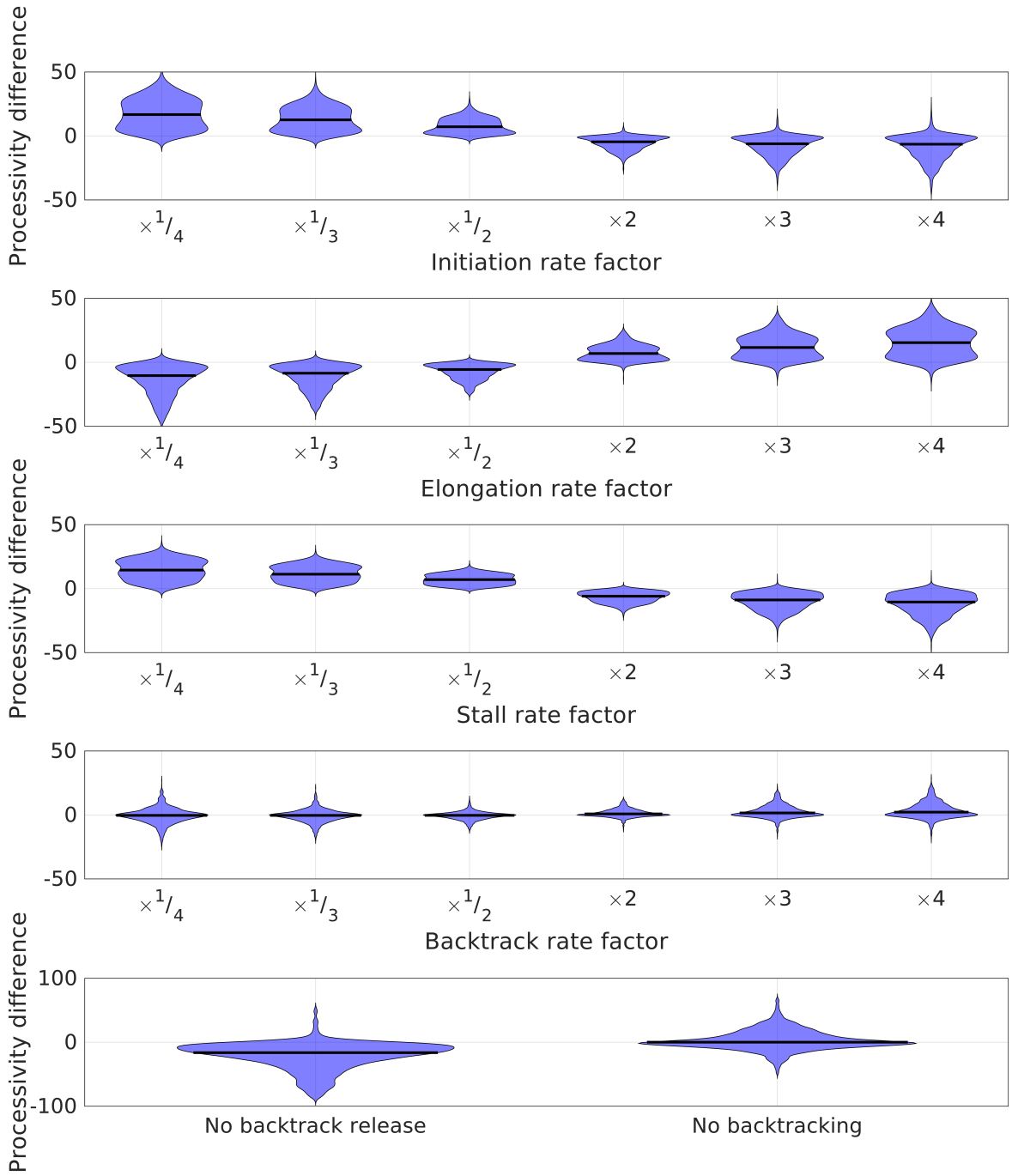


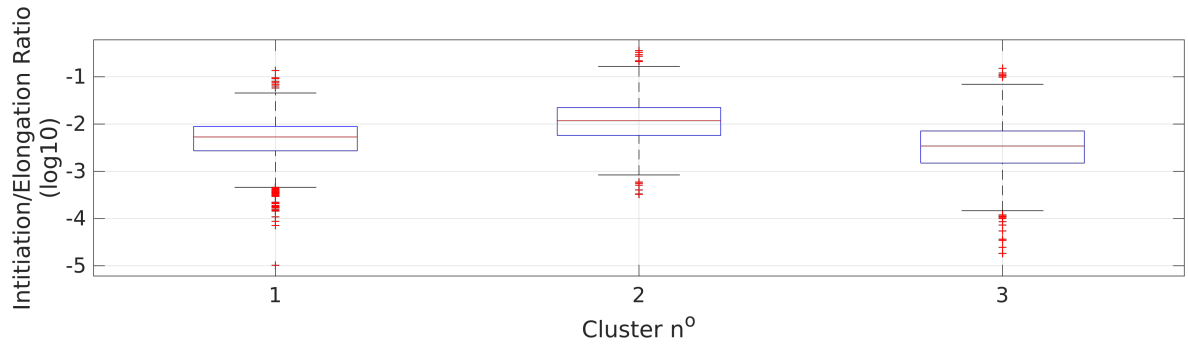
Figure 5.13: Violin plots showing the difference in processivity observed by varying a single parameter in a simulation of transcription. The distribution median is shown as a black line, the shape of the distribution is determined by the kernel function of a histogram of the data. Shown are the distributions of the percentage point changes in processivity as determined by counting number of transcription events that result in a polymerase that reaches the end of the gene, rather than undergoing early termination. Simulations were performed using the same parameter sets as those from Figures 5.8, 5.10 & 5.12.

tails off; (ii) a cluster with a sharp peak of polymerase at the 5' end of the gene; and (iii) a flat, even profile of polymerase.

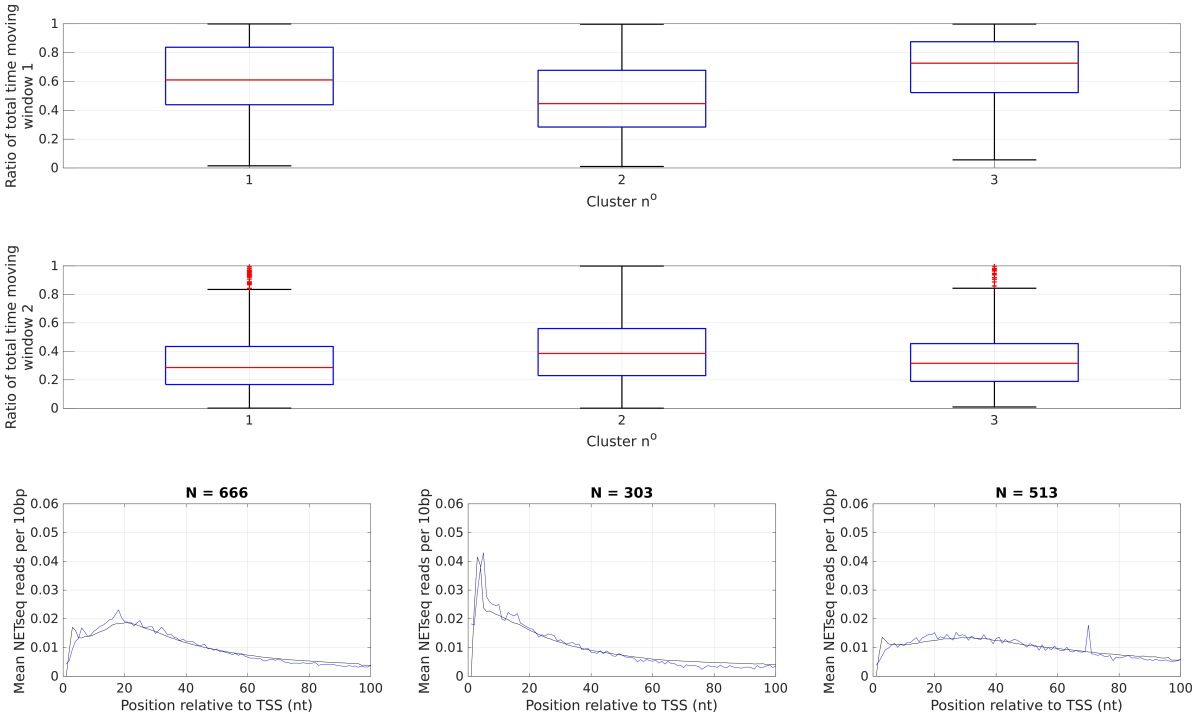
Taking the inferred parameters we find that the sharper profile in cluster 2 is indicative

of a higher ratio of initiation rate to elongation rate than in either of the other two clusters (Figure 5.14a). That this cluster separates from the other two in this manner suggests a key role in the clearing of the promoter region, which could be inhibited by either having elongation occurring too slowly or initiation events too frequently. This cluster with a more pronounced 5' signal also shows a lower moving ratio in the 5' region of the gene and the reverse observation in the second, distal window (Figure 5.14b). This cluster appears to stand out as a group of genes with relatively high initiation rates, a slow promoter-proximal region of polymerase elongation followed by more efficient elongation downstream of some checkpoint, comparable to observations made in *S. pombe* and mouse cells which may be due to the activity of Spt5 (DSIF) (Booth *et al.* 2018, Fitz *et al.* 2018). A experiment such as TEF-seq tracking the locations of polymerase-bound elongation factors may highlight the enrichment or depletion of Spt4/5 or their phosphorylated forms at subsets of genes across the genome, similar to coding and non-coding transcripts respectively enriched and depleted for Paf1 in *S. cerevisiae* (Fischl *et al.* 2017). The third cluster with a more even distribution throughout the gene body stands out as the group of genes with the highest processivity (Figure 5.14c). That all genes appear to have relatively low levels of processivity argues against the idea that transcription has developed as the most efficent method of converting DNA into RNA through a complex environment as in Li & Zhang 2015, yet is in accordance with estimates of transcription processivity observed through *in-vivo* measurements of RNA PolII activity (Steurer *et al.* 2018). The third cluster also stands out as the cluster with lowest initiation-to-elongation ratio and most time moving in the proximal region of the gene. This suggests that promoter escape is a crucial factor in polymerase processivity in *S. cerevisiae*.

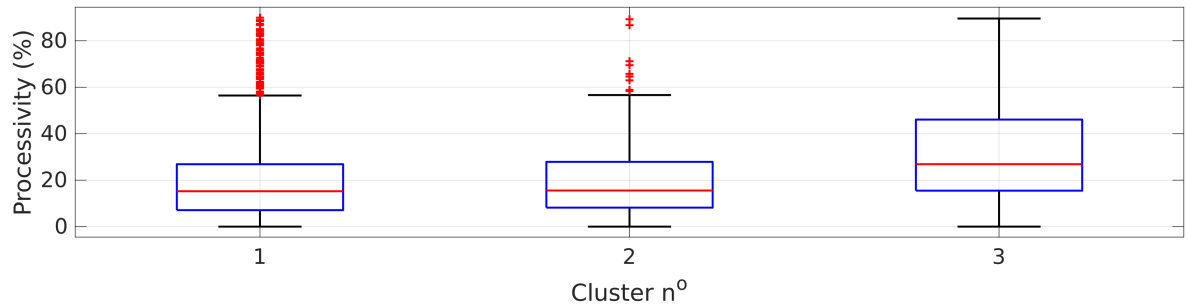
While the presence of two dynamic windows and a centre of early termination are required for the best-fit model, these parameters are not vastly different between the analysed clusters. The size of the first dynamic window is smaller in the first cluster than the third cluster, which displays a more even spread of polymerase, however the second cluster with a more pronounced 5' peak of polymerase does not have a clear size of first window. Furthermore, the locations of early termination events are spread throughout the gene, with some exhibiting promoter-proximal termination and others with termination



(a)



(b)



(c)

Figure 5.14: The parameters which recreate NET-seq distributions are distinct. Shown are the ratios of initiation to elongation rate (5.14a), the moving ratios in the two dynamic windows (5.14b) and the processivity of the parameter sets (5.14c) for the three clusters as defined in Figure (4.3b). Metagene profiles of the clusters as well as fitted distributions are shown in (5.14b).

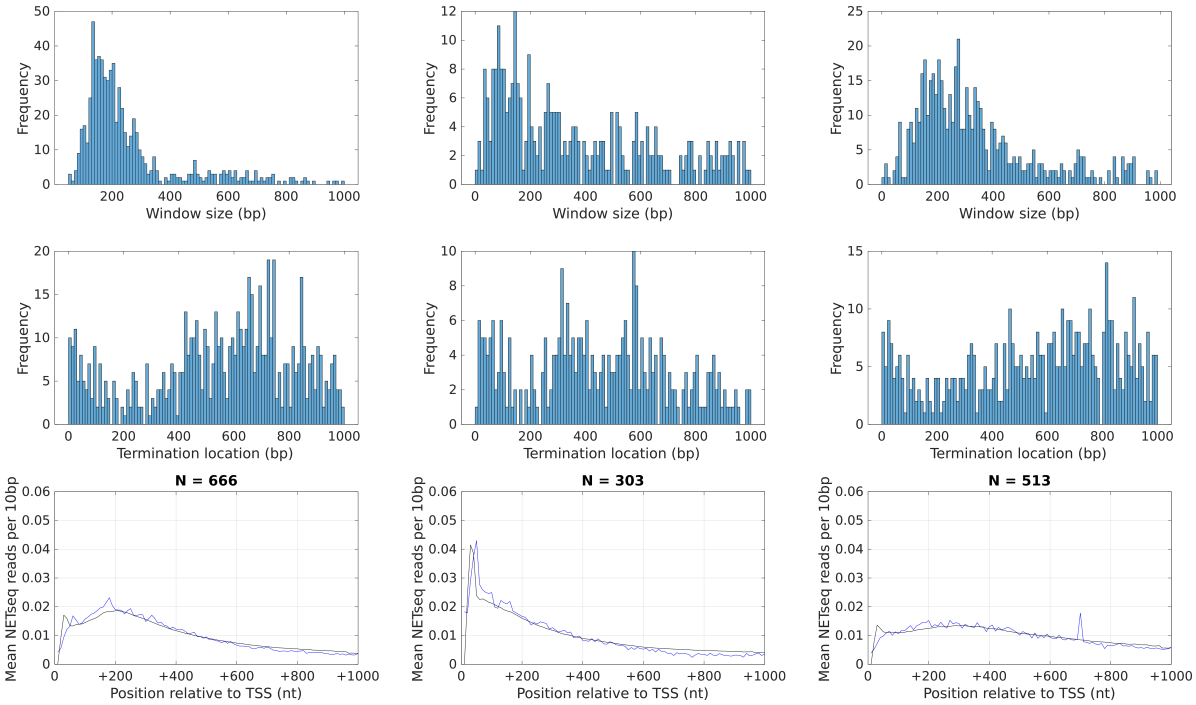


Figure 5.15: The size of the first dynamic window and the central location of early termination as inferred from simulations of wild-type *S. cerevisiae* NET-seq. Histograms showing the locations of these positions for all genes in the clusters as determined in Figure 4.3b. Metagenes of the raw data and fitted profiles for each cluster are shown at the bottom.

in the gene body. This may point to early termination occurring at blocks to transcription which are specific to genes, rather than being a consistent feature of the elongation process where the RNA PolII complex may undergo a decision after a fixed location to either abort or continue elongation.

5.7 Using modelling to obtain numerical transcription rates

The above approach has allowed us to distinguish differences in dynamics that are present at different genes in *S. cerevisiae*. By modelling the profiles of transcriptionally-engaged polymerase at each gene, we are able to separate the key dynamic factors that produce the ranges of distributions of polymerase, however we are still in the dark about what contributes to these dynamics. In this section we will further build on work in Chapter 4, taking the altered profiles of polymerase after genetic intervention and relating these to the dynamic parameters we are able to infer using our modelling approach. By observing how

each parameter varies upon deletion of factors associated with the transcription process, we can construct a network of how these factors function in establishing the kinetics seen *in vivo*.

To analyse the changes in parameter values, we determine significant differences not only by testing significance through calculation of a p-value, but also by determining the magnitude of the differences. We calculate the magnitude of changes in distributions by calculating Cohen's d for the differences in each parameter set (Cohen 1988). The value of Cohen's d is calculated by dividing the mean of the pairwise differences of the parameters at each gene by the standard deviation of these differences:

$$d = \frac{\mu_D}{\sigma_D}$$

The magnitude of the differences is then determined heuristically by the following threshold values given in Cohen 1988:

d>	Difference
0.2	Small
0.5	Medium
0.8	Large

5.7.1 Spt4 assists polymerase processivity

We observed above that the dynamics of transcription affect the proportion of "successful" initiation events at a gene, termed the processivity of a gene. Here, we modelled the NET-seq profiles of genes expressed in both wild-type and mutant strains of *S. cerevisiae*. We found that the changes in parameters from wild-type to mutant at each gene affects of Spt4 on the processivity of each gene (Figure 5.16c).

Deletion of the gene *SPT4* saw many genes become less processive, with a higher number of early termination events seen in the parameter regimes which reproduced these polymerase distributions. The reported role for Spt4 in releasing PolII from proximal pauses in *S. cerevisiae* (Booth *et al.* 2016) could point either to a higher probability of collisions upon deletion of *SPT4* or it may also be the case that the longer residence time on genes of a polymerase undergoing more pausing increases the total probability of a

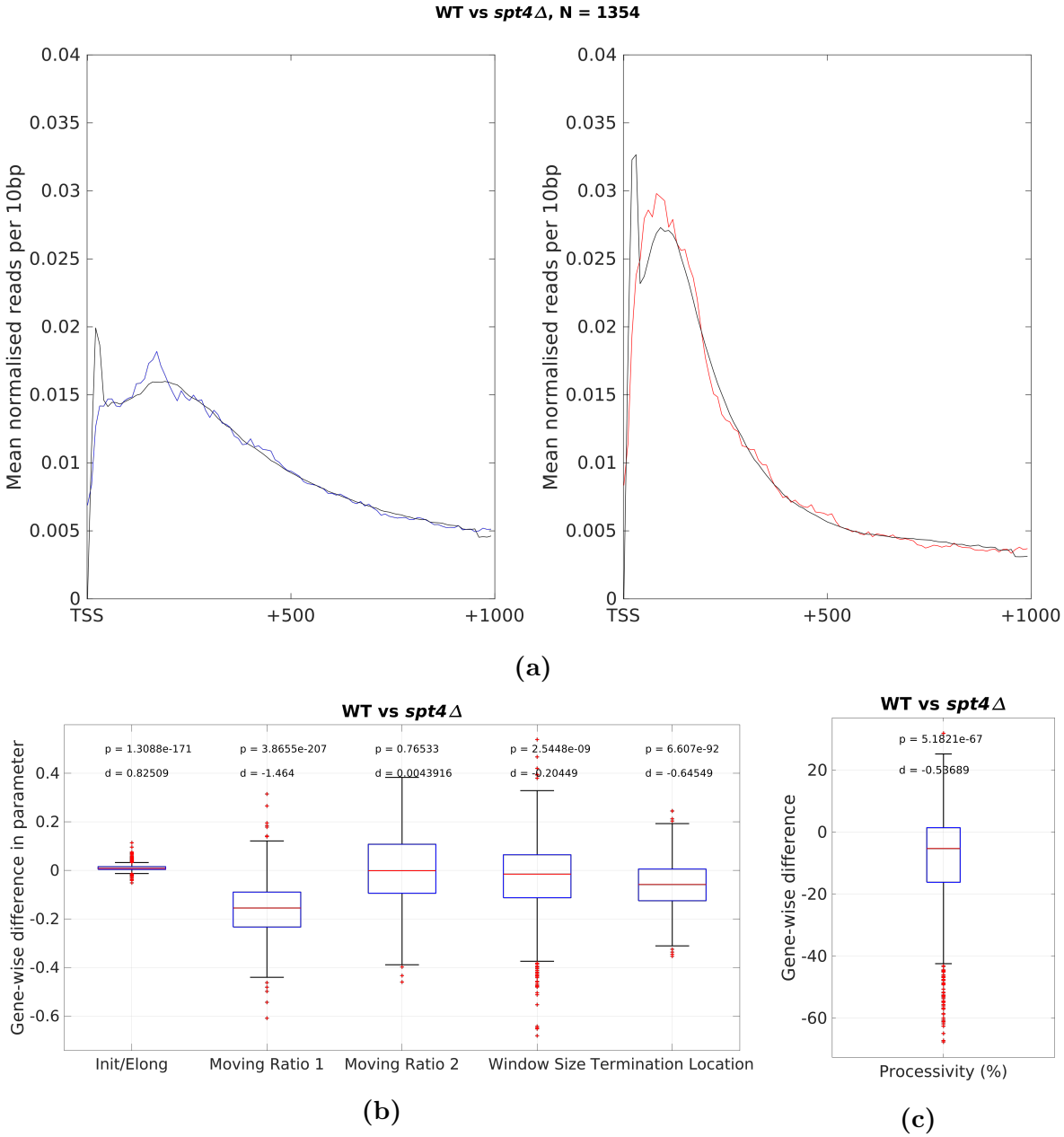


Figure 5.16: Dynamic parameters inferred from fitting model to wild-type and *spt4*Δ *S. cerevisiae* cells. 5.16a shows the raw data for wild-type (blue) and *spt4*Δ (red) NET-seq distributions with corresponding fits in black. 5.16b shows boxplots of the pair-wise differences in inferred parameters for each gene. 5.16c shows boxplots of the difference in processivity for each gene after deletion of *SPT4*. p-value from two-sided sign test, d is Cohen's d for effect size (Cohen 1988). Wild-type data taken from Fischl *et al.* 2017, *spt4*Δ experiment performed by Ülkü Uzun.

polymerase undergoing stochastic early termination during elongation.

The parameter sets which recreated those distributions seen after deleting *SPT4* came about through a reduction in the mean time the polymerase spent moving in the region proximal to the promoter, with no change to this value in the gene body. This was also accompanied by an upstream shift in the mean location of early termination events, occurring closer to the promoter in the *spt4*Δ strain (Figure 5.16b). This suggests a

role for Spt4 in maintaining the contact between the polymerase and the DNA further into the gene body. That the parameter sets also exhibited an increase in the ratio of initiation to elongation rate suggests that without Spt4, polymerase may struggle to clear the promoter region, resulting in more promoter-proximal collisions and further reducing the processivity of transcription events, as was speculated in Hartzog *et al.* 1998.

These results certainly fall in line with Spt4's reported role in a mammalian system as a negative elongation factor at the beginning of genes as part of the DSIF complex with NELF (Kwak & Lis 2013). Spt4 however does exist in this complex with the essential factor Spt5, which undergoes post-translational modification via phosphorylation of its C-terminal domain by Cdk9 (Bur1 in *S. cerevisiae*). This action is clearly not sufficient for full release of a proximal-paused state in *S. cerevisiae* without Spt4. We do not know if after deletion of Spt4, the remaining DSIF subunit Spt5 continues to engage with the polymerase in the same fashion, or if it will then undergo phosphorylation by Bur1. Clearly the full DSIF complex is required for efficient transcription of genes and the small subunit Spt4 has a large role in the dynamics of polymerase from the start to the end of the gene.

5.7.2 Dst1 determines polymerase dynamics in the proximal window

In Chapter 4 we showed that deletion of *DST1* showed a consistent genome-wide shift to NET-seq profiles with a higher proximal/distal ratio in NET-seq reads. While such changes in profiles have been associated with pausing dynamics in the promoter-proximal window, we wished to test this more rigorously, looking not just at pausing dynamics, but the entire elongation process.

Modelling the NET-seq distributions following deletion of *DST1* reveals a role for this annotated backtrack release factor in the promoter-proximal region of genes. The inferred parameters reveal a much less mobile polymerase in the first dynamic window of the genes, while the other transcription parameters are not changed to a large degree (Figure 5.17b). This suggests a role for Dst1 in a fixed window of the genes analysed, yet does not form part of the machinery that determines the locations of any transition

points or checkpoints. As Dst1 has been previously published as the protein responsible for releasing the elongation complex from backtracking events (Churchman & Weissman 2011, Saponaro *et al.* 2014), these results suggest that this activity is either confined to the 5' ends of genes or backtracking events are inherently more likely to occur immediately following transcription initiation.

Potentially as a result of the loss of release from backtracking events, the *dst1*Δ strain exhibits lower processivity at the majority of genes than in the wild-type strain (Figure 5.17c). It has been previously published that strains lacking TFIIS do have an increase in early termination events (Mason & Struhl 2005), here we would suggest that this is a well-conserved result across the yeast genome.

5.7.3 Paf1 is a pause-inducing factor proximal to the promoter

Upon removal of Paf1, we observed a more even distribution of polymerase across the genes analysed. This result is of particular interest given the two contrasting interpretations given for Paf1 in human cells as a maintainer of the promoter-proximal paused polymerase (Chen *et al.* 2015) and conversely a factor to release the paused polymerase (Yu *et al.* 2015).

Here, we observe that deletion of Paf1 decreases initiation/elongation ratio and increases the proportion of time the polymerase spends moving in the promoter-proximal region (Figure 5.18b). That this occurs without an effect on the dynamics in the gene body and without changing the size of the first dynamic window suggests that the role of Paf1 is restricted to the start of the gene body, maintaining a more paused state near the promoter.

This indeed points to Paf1 as a protein which maintains the paused state of the polymerase in the promoter-proximal region, most likely to ensure the correct recruitment of elongation factors and CTD modifications before the polymerase proceeds into the gene body. Paf1 has been shown to join the RNA PolII complex roughly 400bp downstream of the TSS and then remain on the elongation complex until termination (Fischl *et al.* 2017). This suggests that although the effects of Paf1 are seen at the beginning of the gene (Figure 5.18, Chen *et al.* 2015, Yu *et al.* 2015), the recruitment of Paf1 may form a

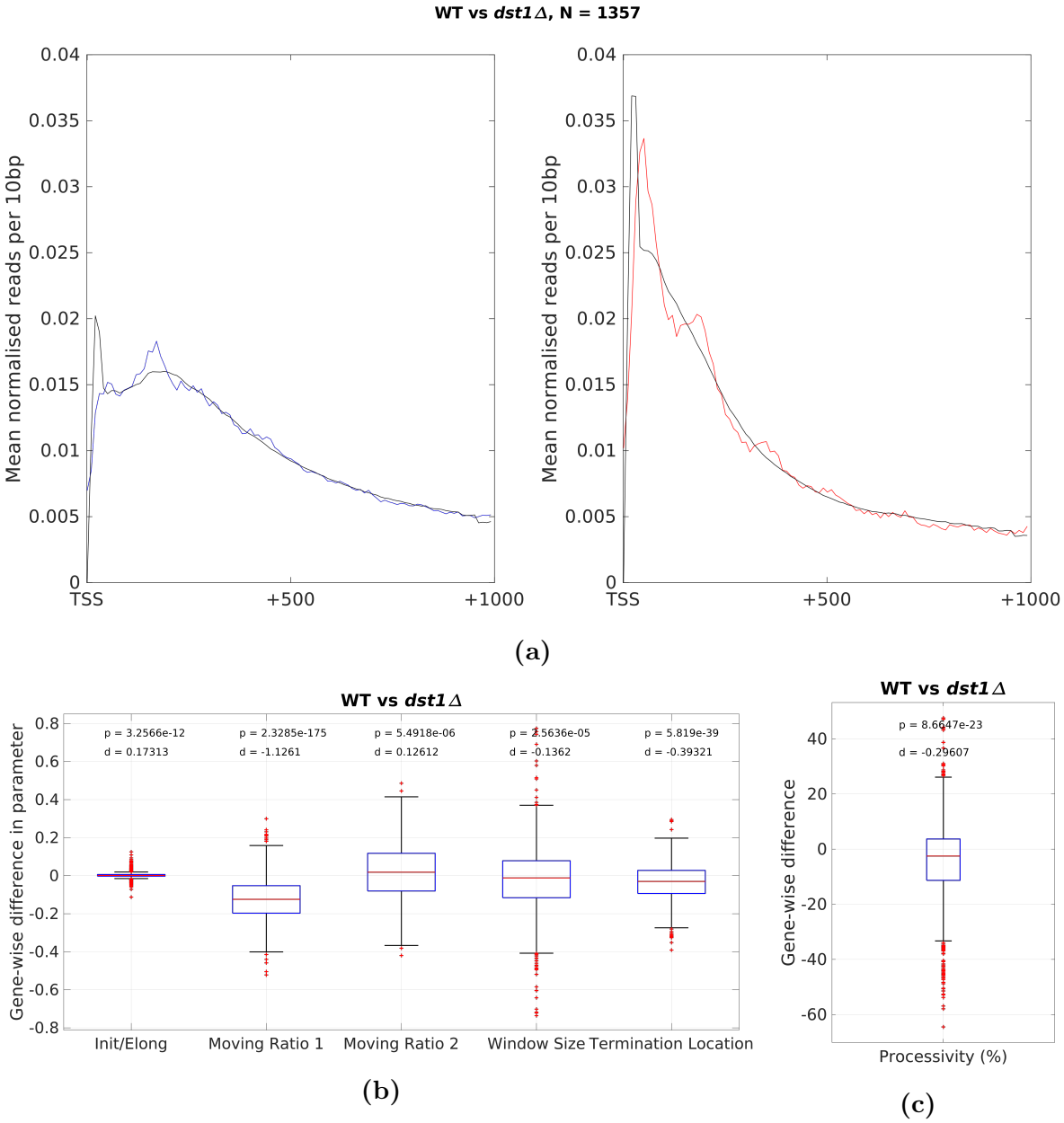


Figure 5.17: Dynamic parameters inferred from fitting model to wild-type and *dst1*Δ *S. cerevisiae* cells. 5.17a shows the raw data for wild-type (blue) and *dst1*Δ (red) NET-seq distributions with corresponding fits in black. 5.17b shows boxplots of the pair-wise differences in inferred parameters for each gene. 5.17c shows boxplots of the difference in processivity for each gene after deletion of *DST1*. p-value from two-sided sign test, d is Cohen's d for effect size. Wild-type data taken from Fischl *et al.* 2017, *dst1*Δ experiment performed by Ülkü Uzun.

crucial role in the subsequent activity of factors involved in pause-release and the clearing of the promoter regions of genes. This efficient release of polymerase from the start of the gene results in more processive transcription at the genes analysed (Figure 5.18c), which is in accordance with the observed increase in RNA seen in Chen *et al.* 2015.

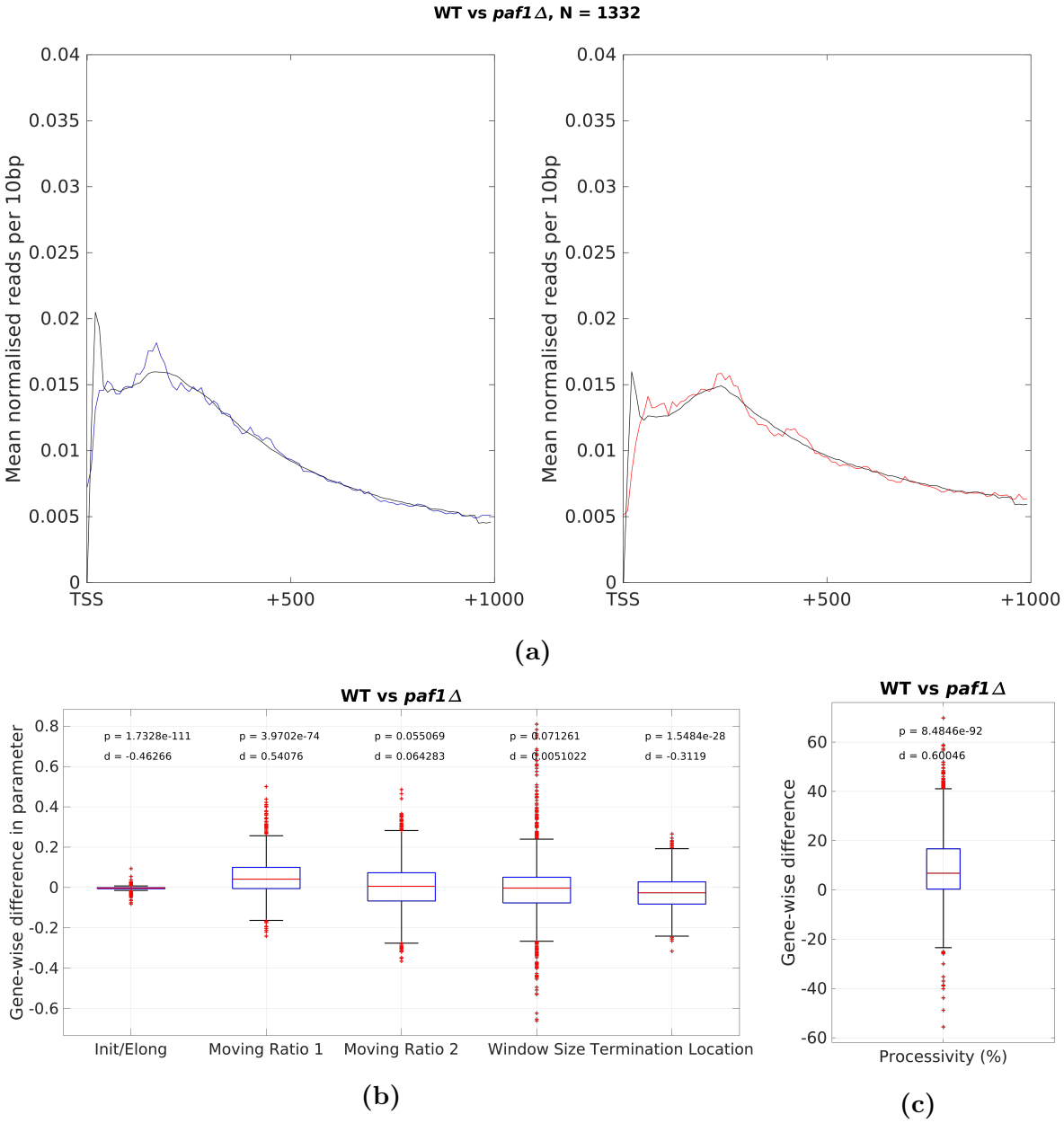


Figure 5.18: Dynamic parameters inferred from fitting model to wild-type and *paf1*Δ *S. cerevisiae* cells. 5.18a shows the raw data for wild-type (blue) and *paf1*Δ (red) NET-seq distributions with corresponding fits in black. 5.18b shows boxplots of the pair-wise differences in inferred parameters for each gene. 5.18c shows boxplots of the difference in processivity for each gene after deletion of *PAF1*. p-value from two-sided sign test, d is Cohen's d for effect size. Data taken from Fischl *et al.* 2017.

5.7.4 Xrn1 inhibits transcription processivity

Similarly to the *paf1*Δ NET-seq data, the distribution of polymerase across genes after deletion of Xrn1 is much more even, with a loss in the 5' peak of polymerase (Figure 5.19a). That deletion of Xrn1 should affect transcription dynamics at all may be a little surprising given its annotated role as a cytoplasmic exonuclease, however this further highlights the fact that not all effects on transcription need be direct. Xrn1 is known to

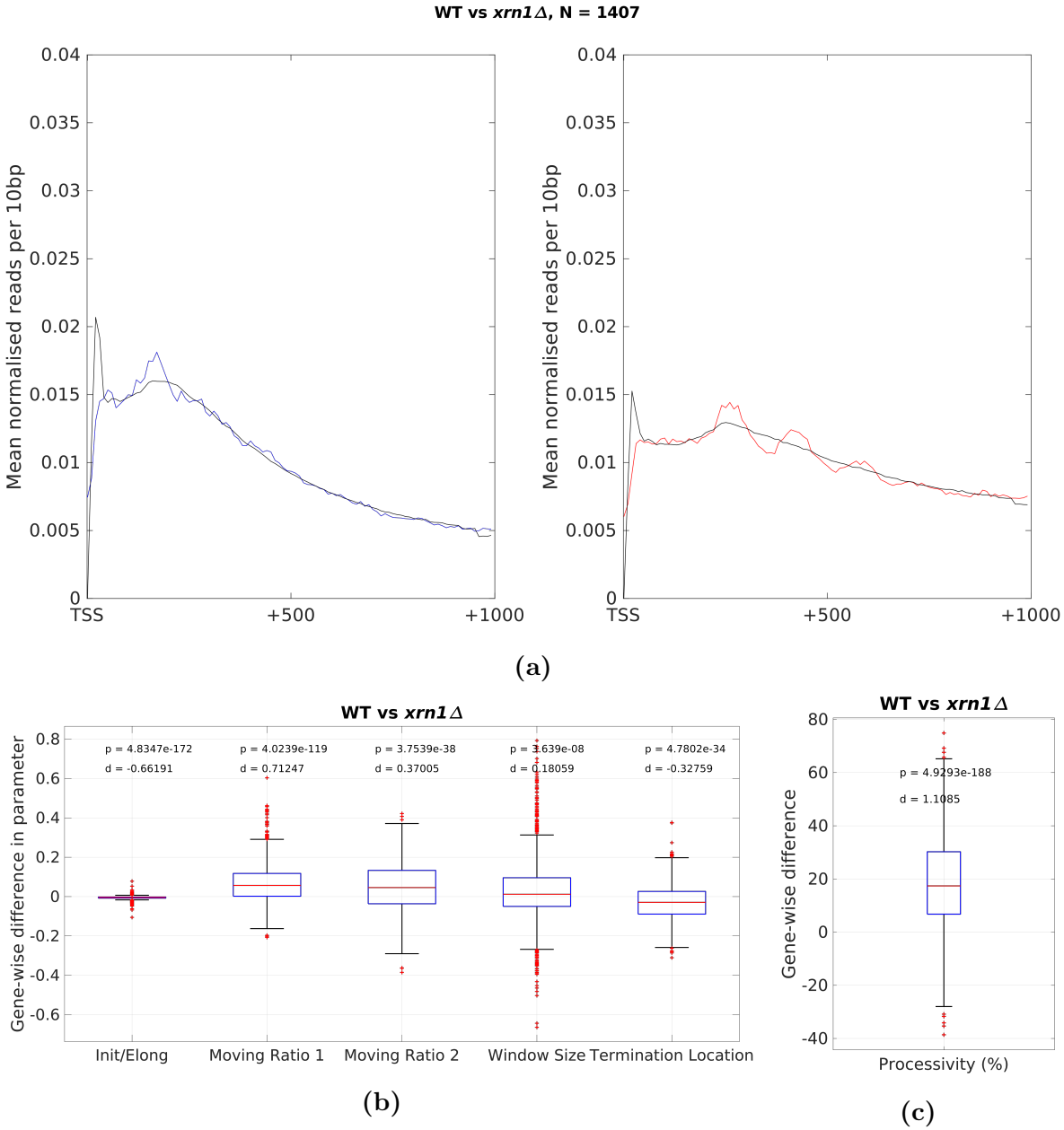


Figure 5.19: Dynamic parameters inferred from fitting model to wild-type and *xrn1*Δ *S. cerevisiae* cells. 5.19a shows the raw data for wild-type (blue) and *xrn1*Δ (red) NET-seq distributions with corresponding fits in black. 5.19b shows boxplots of the pair-wise differences in inferred parameters for each gene. 5.19c shows boxplots of the difference in processivity for each gene after deletion of *XRN1*. p-value from two-sided sign test, d is Cohen's d for effect size. Wild-type data taken from Fischl *et al.* 2017, *xrn1*Δ experiment performed by Harry Fischl.

control the balance of RNA molecules in the cell and particularly target certain genes for degradation, therefore disrupting this balance may have further effects on the concentrations of transcription and elongation factors in the cell.

The deletion of Xrn1 triggers a decrease in the ratio of initiation to elongation rates, potentially allowing for clearing of the promoter region and a consistent increase of the average amount of time the polymerase spends in a moving state (Figure 5.19b). In both

dynamic windows, deletion of Xrn1 increases in the ratio of time moving as opposed to stalled or backtracked, suggesting this affects either the polymerase or the associated factors to the elongation complex. This results in a marked increase in the processivity of the polymerase in the *xrn1*Δ strain, with almost all genes exhibiting at least some increase in processivity (Figure 5.19c).

That Xrn1 has such an important role in the dynamics and processivity of transcription indicates important roles for the degradation machinery in gene expression as a whole, not merely as a control of the RNA turnover. Haimovich *et al.* 2013 suggested, slightly controversially, that Xrn1 shuttles between the nucleus and the cytoplasm and further associates with the chromatin, suggesting a fundamental role as a transcription factor as well as part of the degradation machinery. These results certainly point to a role for Xrn1 in the correct establishment of the elongation complex, with all pausing factors engaged, however we cannot determine whether this is via direct or indirect mechanisms and where in the cell this point of control may be. The techniques used to assess expression and turnover of transcripts in Haimovich *et al.* 2013 (cDTA) relied on measuring rate of synthesis of RNA with labelled nucleotides. While an immediate interpretation of faster production of transcripts may be faster initiation, as was the original interpretation, we would suggest here that this increased rate of production is most likely through an increased elongation rate and a higher processivity of transcription in an *xrn1*Δ strain. This would allow Xrn1 to govern the rates of RNA synthesis indirectly, through maintaining levels of NTPs or the levels of RNA of relevant elongation factors.

5.7.5 Tyrosine 1 of the RNA PolII CTD has a role in promoter-proximal dynamics

The mutation of the Y_1 residue to a phenylalanine (F) in the RNA PolII CTD creates a polymerase complex which cannot be phosphorylated, or dephosphorylated, while undergoing transcription. While a similar mutation in human cells of a portion of the human CTD undergoing the same mutation saw a loss in 5'-accumulation of polymerase and transcriptional read-through (Shah *et al.* 2018), the patterns of Y_1 phosphoylation are different between mammalian cells and *S. cerevisiae* (Zaborowska *et al.* 2016).

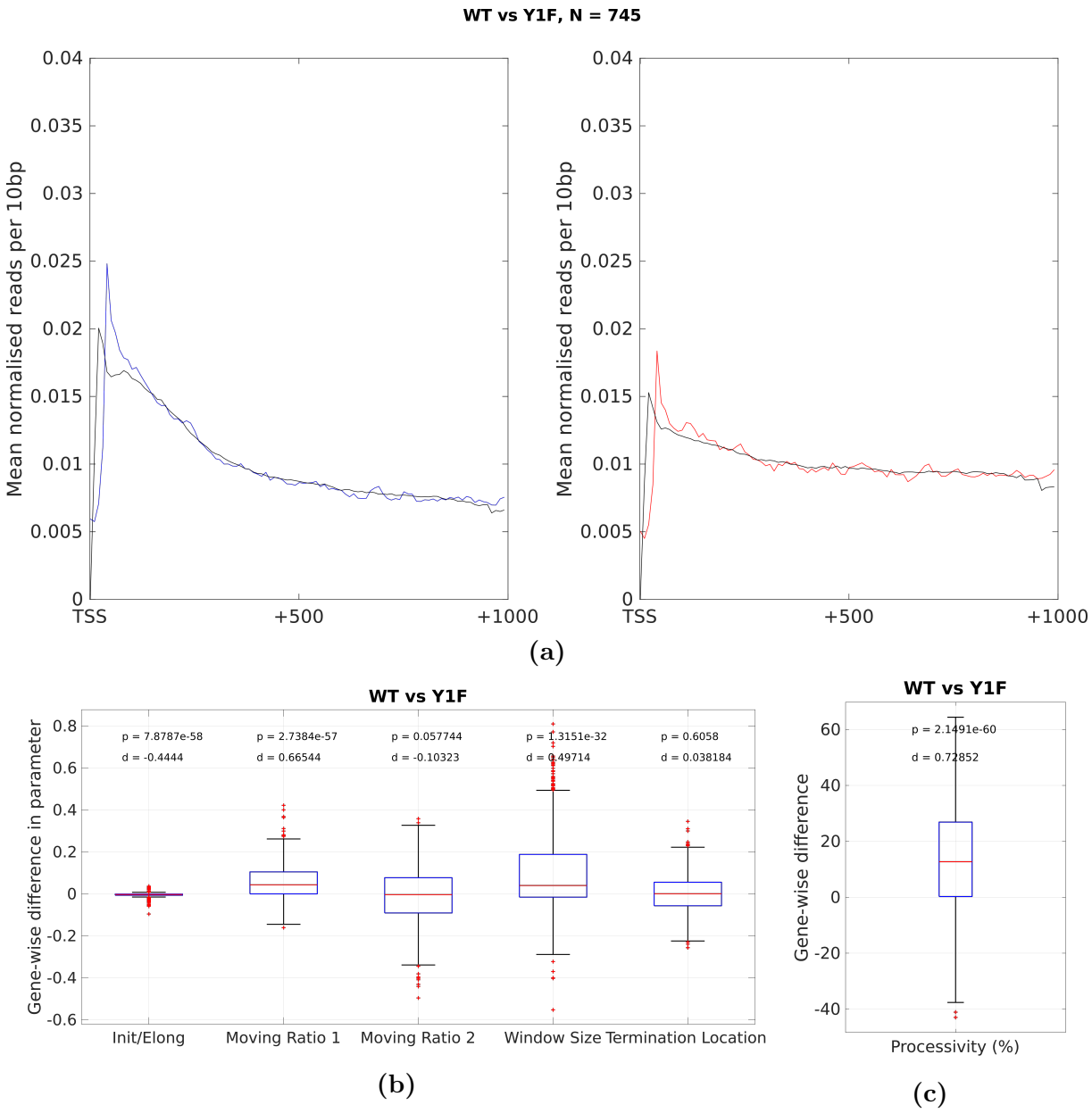


Figure 5.20: Dynamic parameters inferred from fitting model to wild-type and *S. cerevisiae* cells with the Tyrosine in position 1 of the CTD mutated to a phenylalanine (termed Y_1F). 5.20a shows the raw data for wild-type (blue) and Y_1F (red) NET-seq distributions with corresponding fits in black. 5.20b shows boxplots of the pair-wise differences in inferred parameters for each gene. 5.20c shows boxplots of the difference in processivity for each gene after mutation of Y_1. p-value from two-sided sign test, d is Cohen's d for effect size. Data kindly provided by the lab of François Robert (Collin *et al.* 2019).

By modelling NET-seq distributions of actively transcribing polymerase in wild-type and Y_1F cells, we observe a similar drop in the proportion of polymerase accumulating at the 5' end of genes relative to the gene body (Figure 5.20). The changes in parameters which reproduced the distributions observed came about through a decrease in the ratio of initiation to elongation ratio, an increase in the average time spent moving in the proximal window and a larger proximal dynamic window (Figure 5.20b).

We furthermore observed that the parameter sets which recreated the distributions of the Y_1F strain exhibited much more processive transcription, with the majority of genes exhibiting at least a 10 percentage-point increase in processivity under this mutation. These results are in accordance with those recently observed under a similar mutation, where here we infer that the phosphorylation activity of the Y_1 residue serves to keep the polymerase in the proximal region of the gene for longer and it does this by increasing the ratio of time the polymerase spends in a static state. That this first dynamic window is shorter in the wild-type scenario suggests a proximal checkpoint which is governed by the phosphorylation state of the RNA PolII CTD. As the CTD is known to undergo changes around the starts and the ends of genes (Mayer *et al.* 2010, Schlackow *et al.* 2017), this suggests that there may be multiple checkpoints for the polymerase complex around the 5' and 3' ends of the genes, potentially governed by the phosphorylation of the Y_1 CTD residue. This is especially likely given the distribution of Y_1 phosphorylation ChIP'ed downstream of the poly(A) site in *S. cerevisiae* (Figure 1.3, Zaborowska *et al.* 2016). The emerging symmetry of genes in multiple organisms suggests that factors such as Cdk9 (Bur1) may regulate both the stages of initiation and termination in eukaryotes (Parua *et al.* 2018, Laitem *et al.* 2015), therefore it may not be a surprise that the modification of Y_1 at the initial stages of transcription also impacts the stages of termination.

5.7.6 The effects of histone modifications on transcription dynamics

To further test the power of the modelling framework, we wished to establish how deletion of factors involved in the post-translational modifications of histones affect not only NET-seq profiles but the dynamics of the process of transcription. Here, we present the profiles and changes in transcription dynamics upon loss of *RCO1*, a component of the HDAC Rpd3S, involved primarily in the de-acetylation of histones at the 3' end of the gene body (Carrozza *et al.* 2005), Eaf3, part of Rpd3S and the HAT complex NuA4, and the methyltransferase components Set1, Spp1 and Set2.

We found at the *GAL1* locus in Chapter 3 that disrupting the de-acetylase activity of Set3 resulted in an increase of acetylation in the gene body and a decrease in the initiation

and elongation rates of the polymerase. While we may expect the same result here, we must appreciate that if both of these rates decrease by the same factor then there will not be an evident difference in the inferred parameters from the NET-seq simulations. Furthermore, antisense transcription and Set3 have been shown to affect the acetylation levels early in the gene body (Weinberger *et al.* 2012, Murray *et al.* 2015, Brown *et al.* 2018).

Set1, Spp1 and Set2 are all components of the histone methyltransferase machinery. Set1 and Spp1 are part of the COMPASS complex, with Set1 responsible for all methylation marks on H3K4 (me1/2/3) and Spp1 responsible for trimethylation of H3K4, marks associated with active transcription at genes (Howe *et al.* 2017). It is not known how affecting the methylation states of the H3K4 residue affects transcription, or if the marks are a result of transcription itself. Here, we have to tools to investigate how every step of transcription elongation may be affected by loss of either Set1 or Spp1. Set2 places methylation marks on H3K36, which in turn marks genes to undergo de-acetylation by recruiting Rpd3S (Carrozza *et al.* 2005, Venkatesh *et al.* 2012). The interplay between transcription dynamics and post-translational modifications of histones is still currently unknown.

The following NET-seq datasets are taken from Churchman & Weissman 2011 and Howe *et al.* 2017.

Rpd3S components assist polymerase progression through the gene

While Eaf3 appears to have a more pronounced effect on the profiles of polymerase by looking at averaged profiles than Rco1 (Figures 4.11 & 4.12), the change to underlying dynamics from modelling the distributions reveal consistent roles on the dynamics of polymerase, however there still may be compensatory effects through disruption of both acetylation and de-acetylation complexes. Deletion of either subunit was found to reduce moving ratio of the polymerase. The main difference between the two is that Eaf3 appears to control polymerase movement exclusively in the distal region of genes, whereas Rco1 assists polymerase movement throughout the entire gene (Figures 5.21 & 5.22).

This is in agreement with the FISH modelling at the single locus in Chapter 3, where

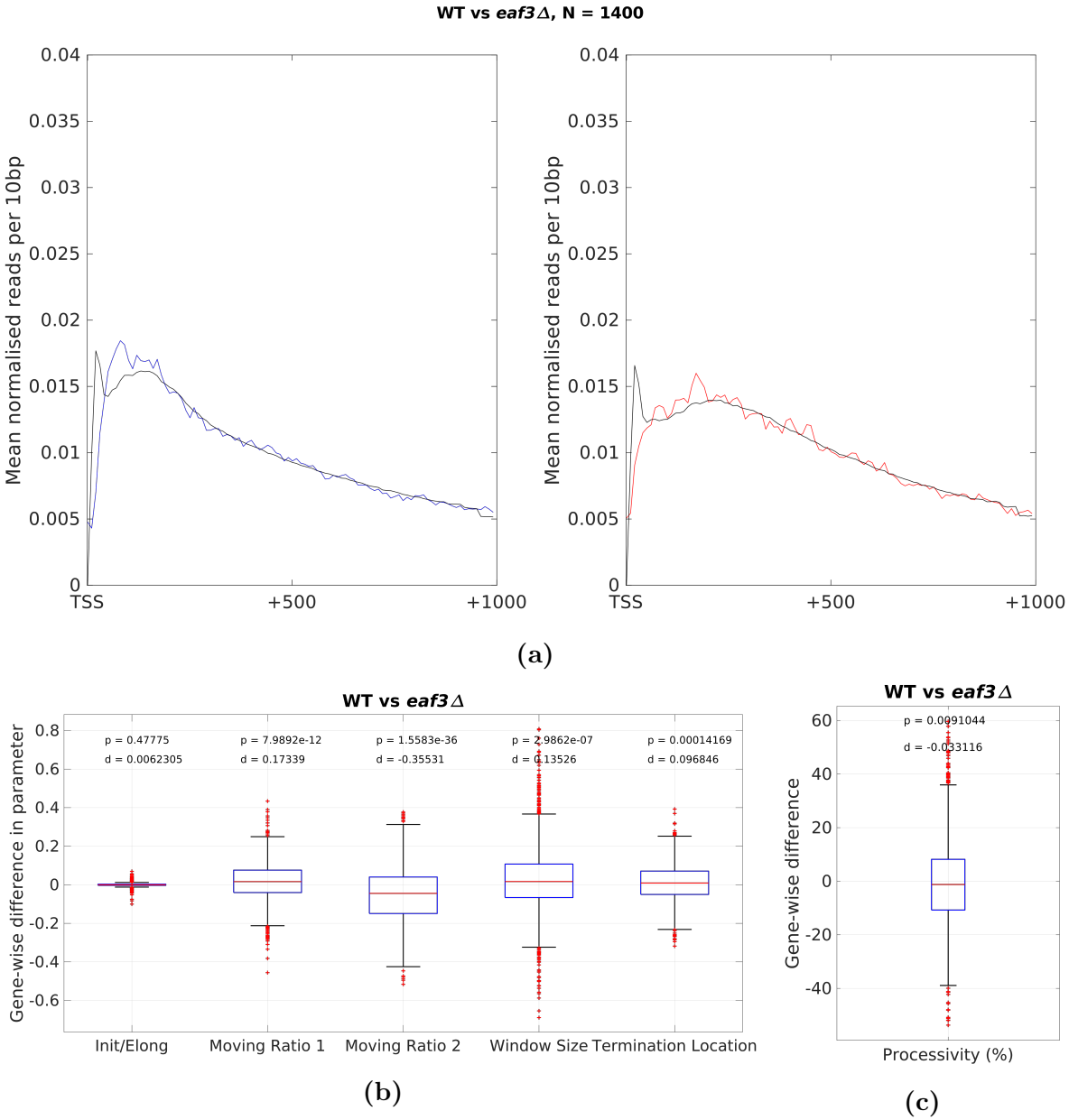


Figure 5.21: Dynamic parameters inferred from fitting model to wild-type and *eaf3*Δ *S. cerevisiae* cells. 5.21a shows the raw data for wild-type (blue) and *eaf3*Δ (red) NET-seq distributions with corresponding fits in black. 5.21b shows boxplots of the pair-wise differences in inferred parameters for each gene. 5.21c shows boxplots of the difference in processivity for each gene after deletion of *EAF3*. p-value from two-sided sign test, d is Cohen's d for effect size. Data is taken from Churchman & Weissman 2011.

we saw that deletion of the histone-deacetylase Set3C subunit Set3 resulted in a reduction in elongation/export rate. Here, by deleting elements of Rpd3S we similarly expect an increase in histone acetylation at these genes and observe a slower polymerase through the gene body. We are able to separate the stalling dynamics of the polymerase from the speed of a polymerase and find that the effect of deleting Eaf3 is on the stalling of polymerase in the gene body, whereas Rco1 affects the stalling throughout the gene

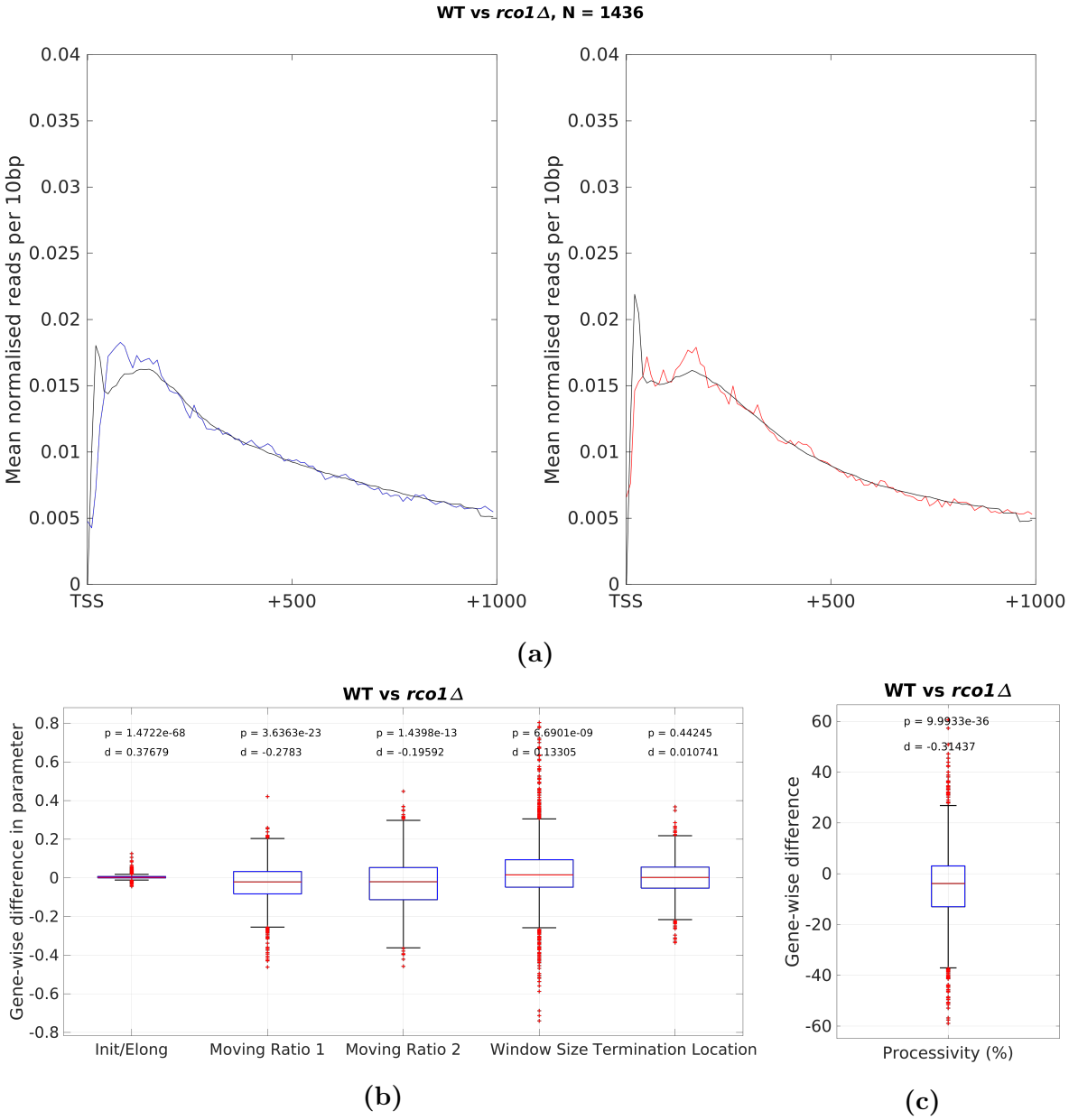


Figure 5.22: Dynamic parameters inferred from fitting model to wild-type and *rco1*Δ *S. cerevisiae* cells. 5.22a shows the raw data for wild-type (blue) and *rco1*Δ (red) NET-seq distributions with corresponding fits in black. 5.22b shows boxplots of the pair-wise differences in inferred parameters for each gene. 5.22c shows boxplots of the difference in processivity for each gene after deletion of *RCO1*. p-value from two-sided sign test, d is Cohen's d for effect size. Data is taken from Churchman & Weissman 2011.

body as well as disrupting the initiation/elongation ratio. This suggests that there are conflicting roles for Eaf3 in establishing the elongation dynamics in the promoter-proximal region of genes, with Rpd3S acting to increase the moving ratio of the polymerase and NuA4 counteracting this by inducing a higher ratio of stalling or backtracking.

As we showed in Chapter 4, the effects of loss of Eaf3 and Rco1 may be confined to subsets of genes. Indeed, looking at the overall distributions of changes in parameters in

Figures 5.21b & 5.22b, there are a large number of genes which do show large changes in these parameters. This is in line with observations from Brown *et al.* 2018 that genes with and without antisense react differently to loss of HDACs, with genes with antisense exhibiting a smaller increase in acetylation upon deletion of *SET3*. While the distinguishing feature was not shown to be antisense transcription for loss of Rpd3S components (Brown *et al.* 2018), there may be other factors that set apart these groups of genes which are more sensitive to loss of Rpd3S activity, impacting the dynamics of transcription. A further study observed that the action of histone acetylation may be to regulate the step from early transcription into elongation at a subset of genes (Gates *et al.* 2017). Similarly to the action of Paf1, the acetylation of histones may act to hold the polymerase in the promoter region, to ensure the relevant elongation factors and modifications are associated with the RNA PolII complex before proceeding into the gene body.

Histone methyltransferases assist polymerase through the gene body

While we did not observe large changes in the genome-averaged NET-seq profiles after deletion of the histone methyltransferases Set1, Spp1 and Set2 in Chapter 4, modelling the individual profiles reveals a more complex picture. Set1 and Spp1 are known to exist in the same complex (Set1C/COMPASS), which targets H3K4 for methylation, however Spp1 is only responsible for the tri-methylation of H3K4, whereas Set1 is responsible for mono-, di- and tri-methylation (Howe *et al.* 2017).

Modelling the NET-seq profiles after removal of these subunits of the Set1C complex reveal different roles for these proteins, and potentially the methylation state of H3K4 on the dynamics of transcription. Set1 is shown to assist the polymerase through the gene body by reducing the stalling and backtracking frequency in both the proximal and distal regions of the gene (Figure 5.23). Loss of Spp1, however, did not see any change in the dynamics of transcription, suggesting distinct roles for these two factors, with Set1 assisting polymerase elongation and Spp1 having no immediately evident effect on transcription dynamics (Figure 5.24).

It furthermore suggests different effects of the different methylation states of H3K4. It is difficult to separate the roles of H3K4 mono- and di-methylation, as Set1 is involved

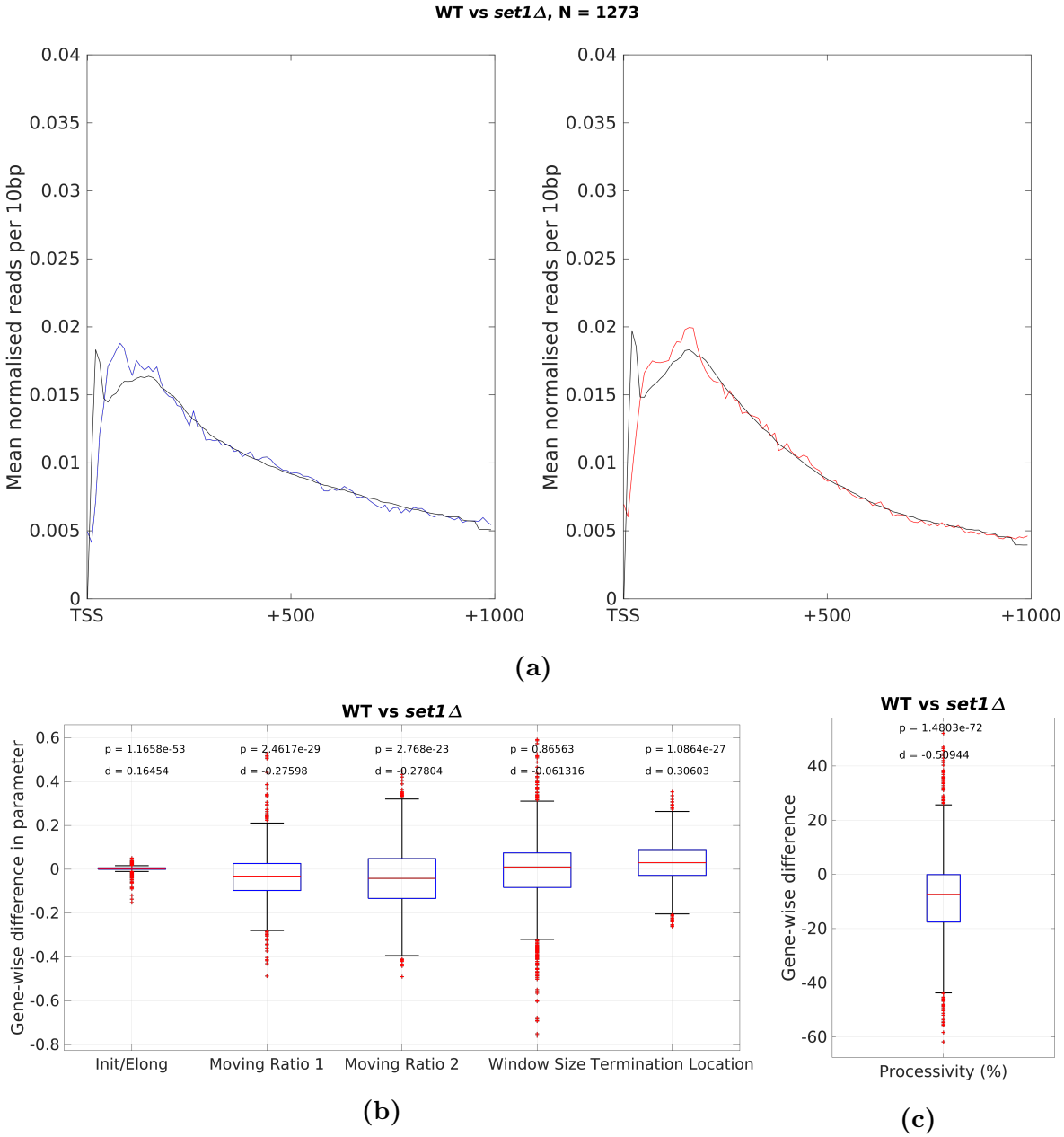


Figure 5.23: Dynamic parameters inferred from fitting model to wild-type and *set1*Δ *S. cerevisiae* cells. 5.23a shows the raw data for wild-type (blue) and *set1*Δ (red) NET-seq distributions with corresponding fits in black. 5.23b shows boxplots of the pair-wise differences in inferred parameters for each gene. 5.23c shows boxplots of the difference in processivity for each gene after deletion of *SET1*. p-value from two-sided sign test, d is Cohen's d for effect size. Data is taken from Churchman & Weissman 2011.

in the placement of both of these marks, however they clearly have distinct roles to that of H3K4me3 when comparing the *set1*Δ and *spp1*Δ transcription rates. As di- and tri-methylation have been shown to accumulate over longer periods of transcriptional activity at individual genes (Soares *et al.* 2017), it may be the case that only H3K4me1 is important in establishing the dynamics of transcription, with the further methylation states responsible for setting up the chromatin environment for other processes such as

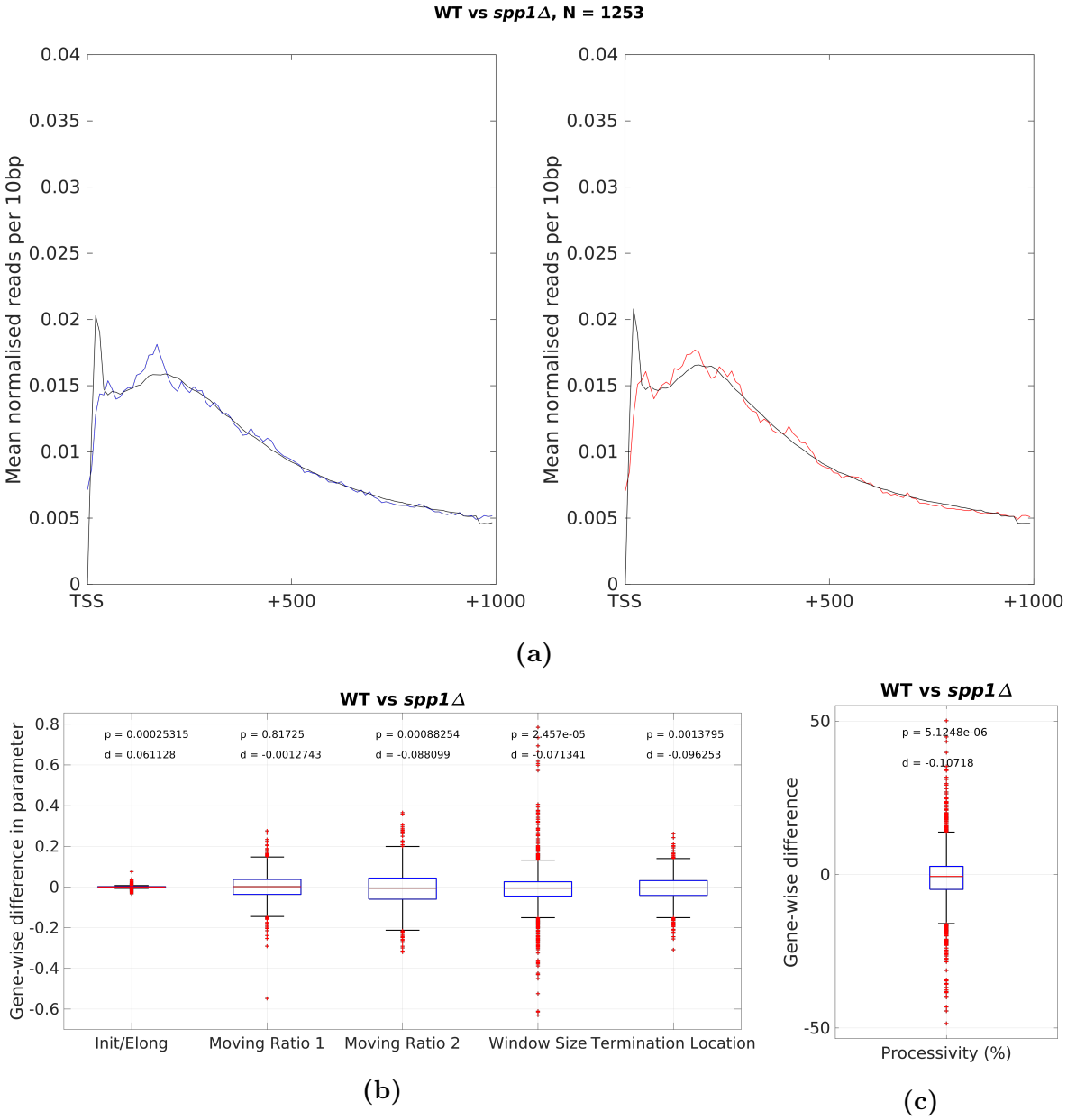


Figure 5.24: Dynamic parameters inferred from fitting model to wild-type and *spp1*Δ *S. cerevisiae* cells. 5.24a shows the raw data for wild-type (blue) and *spp1*Δ (red) NET-seq distributions with corresponding fits in black. 5.24b shows boxplots of the pair-wise differences in inferred parameters for each gene. 5.24c shows boxplots of the difference in processivity for each gene after deletion of *SPP1*. p-value from two-sided sign test, d is Cohen's d for effect size. Data is taken from Howe *et al.* 2017.

chromosome segregation (Karányi *et al.* 2018) or, as has been recently suggested, involved in setting up genes to undergo further rounds of activation under changing external carbon sources, guiding the Transcriptional Repression Memory (TREM, Lee *et al.* 2018).

The removal of Set2 in *S. cerevisiae* resulted in similar effects to the loss of Set1. Set2 is responsible for the methylation of H3K36 which, like H3K4, can be modified with multiple methyl groups. The loss of Set2 resulted in an increase in the proportion of

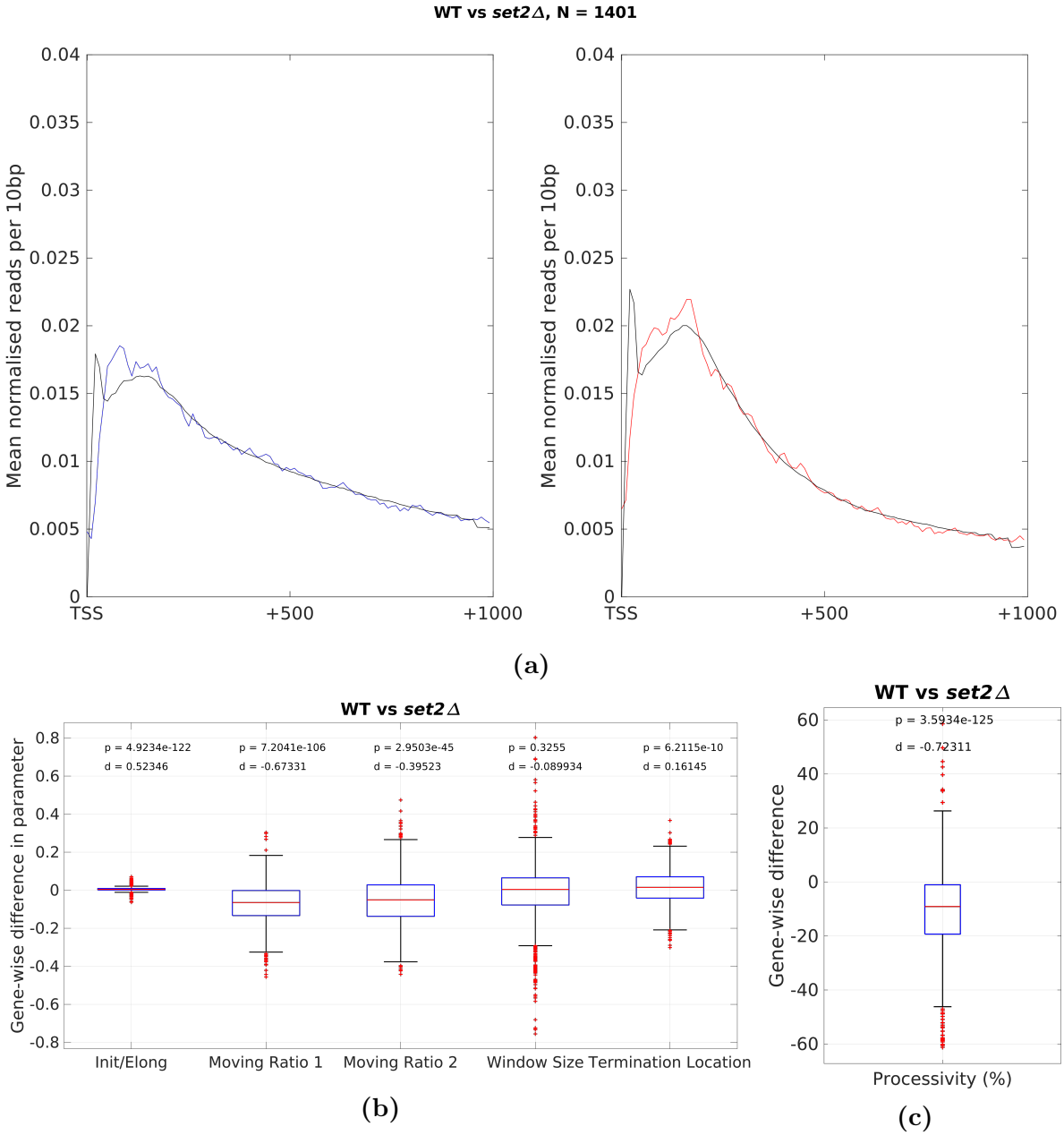


Figure 5.25: Dynamic parameters inferred from fitting model to wild-type and *set2*Δ *S. cerevisiae* cells. 5.25a shows the raw data for wild-type (blue) and *set2*Δ (red) NET-seq distributions with corresponding fits in black. 5.25b shows boxplots of the pair-wise differences in inferred parameters for each gene. 5.25c shows boxplots of the difference in processivity for each gene after deletion of *SET2*. p-value from two-sided sign test, d is Cohen's d for effect size. Data is taken from Churchman & Weissman 2011.

stalling throughout the gene, as with Set1 and an increase in the initiation/elongation ratio, suggesting the activity of Set2 may assist in the clearing of the promoter region of the gene (Figure 5.25). Again, the recruitment of Set2 and the subsequent methylation of H3K36 may be a consequence of transcription itself, feeding back to the elongation complex to assist in further rounds of transcription or Set2 may methylate a non-histone target.

5.8 Transcription dynamics in HeLa cells

Qualitatively, the distributions of transcriptionally-engaged polymerase in HeLa cells appeared to exhibit much more pronounced peaks of polymerase at the 5' ends of genes than in *S. cerevisiae*. However, now we have a modelling setup to analyse the mNET-seq data in higher detail, we can investigate whether the key steps involved in polymerase elongation are different between *S. cerevisiae* and HeLa cells. Furthermore, we are able to determine how the rates of elongation, stalling or backtracking differ between the two organisms and, importantly, how much overlap exists in the regulation of transcription from lower to higher eukaryotes.

5.8.1 Transcription dynamics are conserved between *S. cerevisiae* and HeLa cells

We wished to extend the analysis of transcription beyond the dynamics in the simple eukaryote *S. cerevisiae*. We were able to apply the same modelling setup as above to the comparable technique mammalian-NET-seq (mNET-seq), developed to study the locations of transcriptionally-engaged polymerase in higher eukaryotes and applied to HeLa cells (Nojima *et al.* 2015, Schlackow *et al.* 2017). The human genome shares many aspects with *S. cerevisiae*, however there are a number of key differences that need to be taken into account. Firstly, human genes tend to be much longer than yeast genes, on the order of tens or hundreds of kb, rather than the mean length of 1,500bp in yeast. As such, more genes will fall into our inclusion criteria of being longer than 1,000bp, giving more power to our analysis. Secondly, almost all human genes contain introns, and the length of the first exon is around 500-600 bp in length (Brown *et al.* 2018). This may create problems interpreting the mNET-seq signal across the first 1,000bp, as splicing is known to affect the rate of trancription and vice-versa (Nojima *et al.* 2015, Aslanzadeh *et al.* 2018). However, given the non-uniform lengths of the first exon, any effects, which are often on nucleotide-scale, should be averaged out over the genome in our analysis.

We repeated the maximum-likelihood approximation approach (Section 5.3, as defined above) with the mNET-seq data to determine the best-fit model via calculation of the AIC. We found that the best-fit model for the genes included was the same as that for

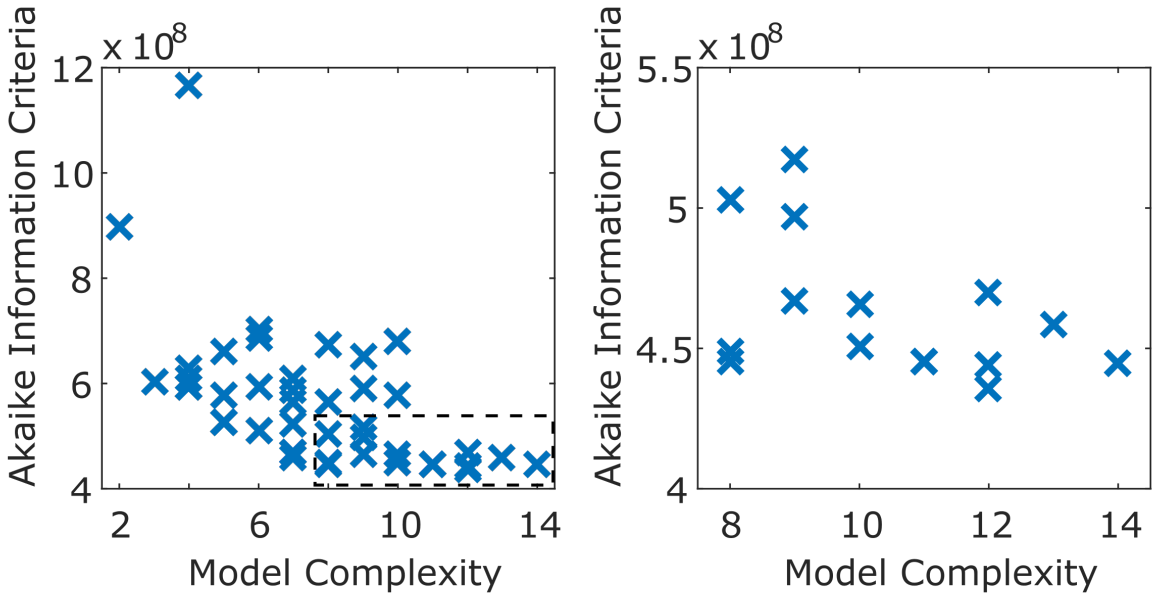


Figure 5.26: Akaike Information Criteria (AIC) for the analysed nascent transcription data obtained for HeLa cells via mNET-seq in Figure (4.16). Right shows an enlargement of the marked area.

S. cerevisiae NET-seq, namely a model with stochastic stalling and backtracking, two windows with different stalling and backtracking parameters, early termination around a mean position and a non-bursty promoter (Figure 5.26).

We furthermore tested how consistent this choice of model was across the the genome, splitting the data into the six clusters defined in Figure 4.16b and repeating the AIC calculation for each sub-group of genes. We found that in 4 out of 6 clusters the same model as fit the entire genome was determined to be the best-fit model (Figure 5.27). However, for two clusters (clusters 3 & 4), which had pronounced peaks of polymerase roughly 50-100bp downstream of the TSS, the best-fit model was a model which included a bursty promoter, variable start-site selection, stochastic stalling and two windows of stalling parameters. This interestingly highlights the fact that some genes may exhibit behaviours typical of bursting promoters as seen in Chapter 3, where the NET-seq distributions across a population of cells do not average out any bursting effects. Variable start-site selection also may be a more important factor in mammalian transcription, as genes are known to have many isoforms which differ in length more than those isoforms identified in *S. cerevisiae* (Floor & Doudna 2016, Pelechano *et al.* 2014).

However, the majority of genes analysed from the human mNET-seq experiment

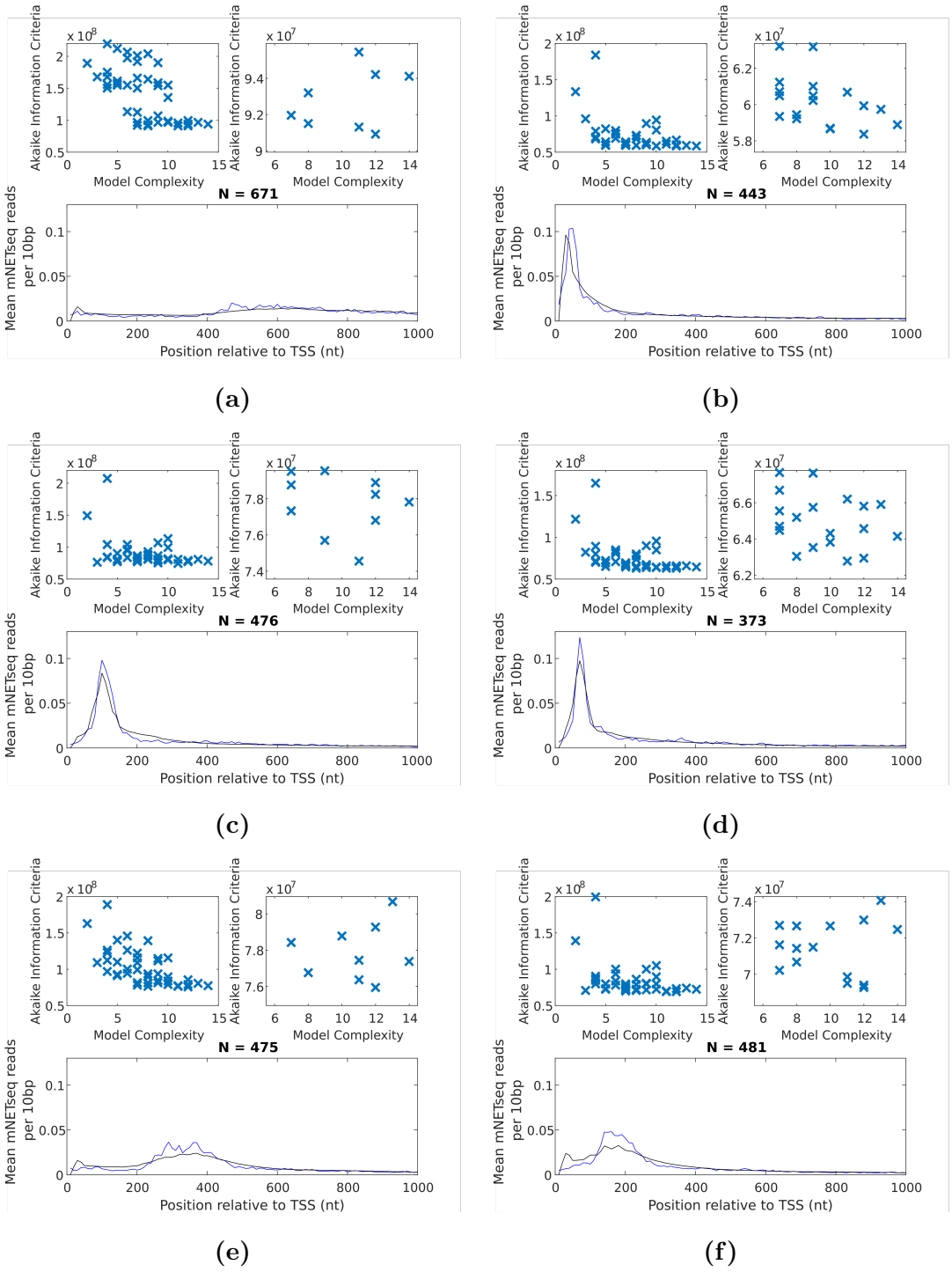


Figure 5.27: Akaike information Criteria (AIC) and metagene plots of best-fit simulated distributions for the six clusters of data as given in Figure 4.16b. Shown are the AIC values for all models and an enlargement of the AIC values for the best models (top-right). Metagene profiles show raw data in blue and fitted distributions in black.

(∼70%) are best described by the model determined as the best-fit for both *S. cerevisiae* and HeLa cells as well as being true when considering a universal model for all genes in either genome. This suggests that transcriptional dynamics are well conserved from lower to higher eukaryotes and that many shared factors are likely to be performing

the same roles in both organisms.

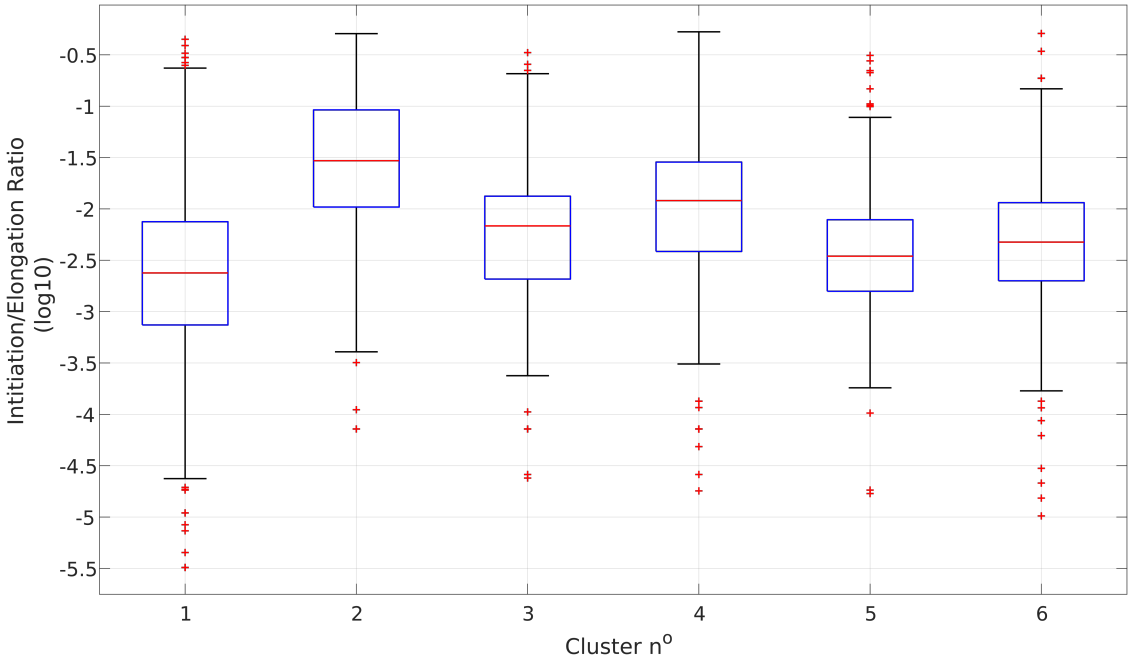
5.8.2 Transcription dynamics are distinct across the human genome

The clusters of mNET-seq profiles observed in Figure 4.16 are visually much more different from each other than the equivalent clusters determined in *S. cerevisiae*. Therefore, we wish to further understand how the dynamics differ at genes which exhibit sharp peaks of polymerase, to those which have much flatter, more even profiles of polymerase. We observed in Chapter 4 that the shape of the polymerase distribution has little impact on the level of mature RNA in the cell, yet this does not give us an indication of the rates of production of the transcripts and the process of polymerase elongation.

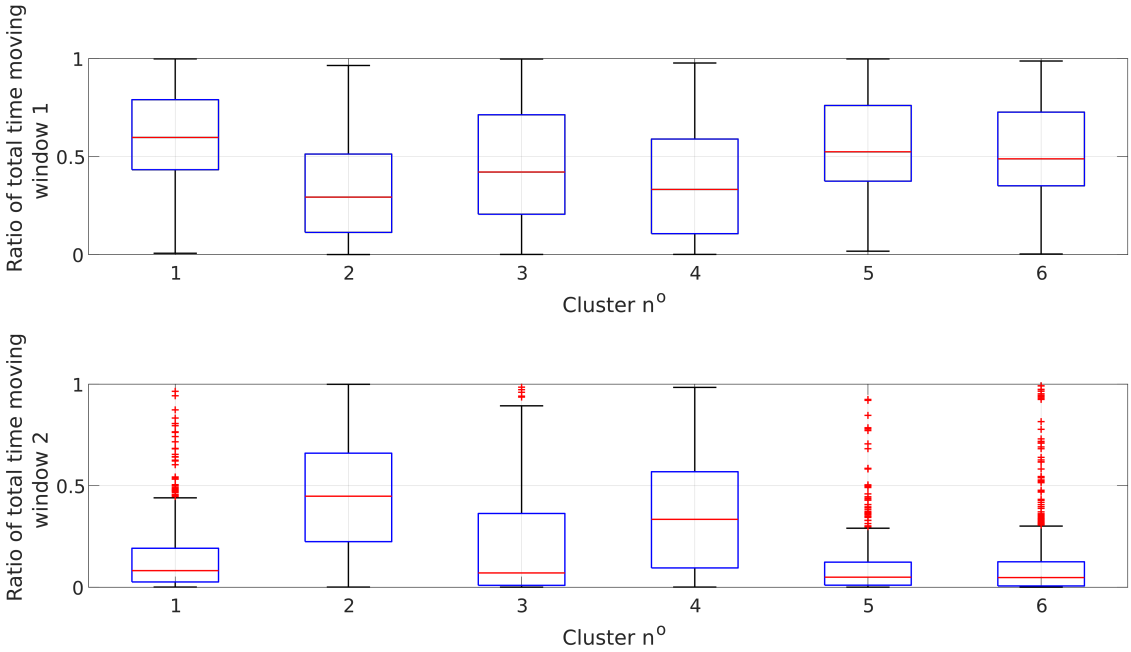
Similarly to those results from the analysis of NET-seq data from *S. cerevisiae*, we split the parameter results into 5 main components. We find that those clusters which exhibit flatter profiles of polymerase across the gene are firstly associated with lower ratios of initiation to elongation than those with more pronounced 5' peaks of polymerase (Figure 5.28). The first result of interest is that there is a lot of overlap between the range of these parameters in the human and *S. cerevisiae* results in Figure 5.14. The main difference between the inferred initiation/elongation ratios in the human system to *S. cerevisiae* is that the clusters are much more different from each other, with less overlap. This is likely to be due to the true kinetics than any statistical artefact as the chosen clusters for the two organisms are of roughly the same numbers of genes.

In terms of the mean time the polymerase spends moving, the separation between clusters appears more pronounced when comparing the moving ratio in the second window than in the first (Figure 5.28b), with the the flatter profiles in clusters 1, 5 & 6 showing much lower moving ratios in the distal window that those with more "peaky" profiles. The flatter profiles also exhibit a higher mean proportion of time with moving polymerases in the first window, but the main comparator that sets the clusters apart is the ratio between the first window and the second. Clusters 1, 5 & 6 have a much higher ratio of moving in the first window than the second, which would reflect the ability of the polymerases at these genes to rapidly clear the promoter region before proceeding into the gene body.

However, these results comparing the moving ratios in the two dynamic windows must



(a)



(b)

Figure 5.28: Inferred parameters from modelling mNET-seq profiles in the 6 clusters determined in Figure 5.27. Shown are the ratios of initiation to elongation rate for each cluster (5.28a) and the mean time spent in a moving state as opposed to stalled or backtracked in the two dynamic windows (5.28b).

also be made in the context of the size of the first dynamic window at each gene. There are much more pronounced locations in each cluster which demarcate the switch from the first to second dynamic regimes in the HeLa cells than in *S. cerevisiae* (Figure 5.29). Those clusters which do exhibit a peak of polymerase near the 5' region of the gene (Clusters 2

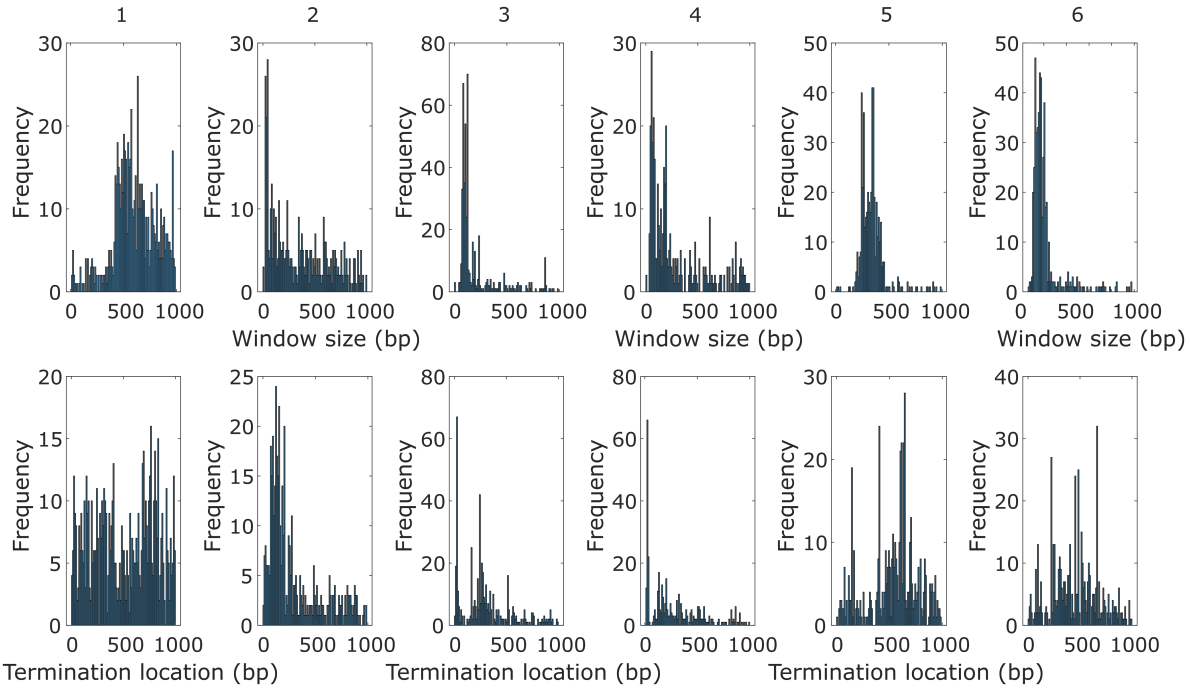


Figure 5.29: Inferred sizes of the first dynamic window and the mean location of early termination for the 6 clusters of mNET-seq profiles determined in Figure 5.27.

- 6) have a pronounced location where the first window ends, varying from 50bp - 400bp downstream of the TSS.

One main difference between the human results and those from *S. cerevisiae* are that those genes with a sharp peak of polymerase close to the TSS have more consistent locations of early termination (Figure 5.29). The flatter profiles have termination locations spread evenly across the gene body, similar to *S. cerevisiae*. This suggests that the genes in Clusters 2, 3 & 4 differ in their regulation of polymerase elongation than *S. cerevisiae* by having clear points of early termination, possibly as a checkpoint to maintain the correct construction of the elongation complex during initiation which acts as a point-of-no-return for the polymerase complex as has been previously suggested (Tellier *et al.* 2016).

Estimates of polymerase processivity are much lower in HeLa cells than in *S. cerevisiae*, which may not be surprising with recent estimates putting a global estimate on polymerase processivity in human cells at 1% using RNA PolII FRAP (Steurer *et al.* 2018). While again, the processivity estimates from Cluster 1 are more similar to those derived from NET-seq in *S. cerevisiae* in the region of 10-40%, in all other clusters the processivity is below 20% or even below 10% for the majority of genes.

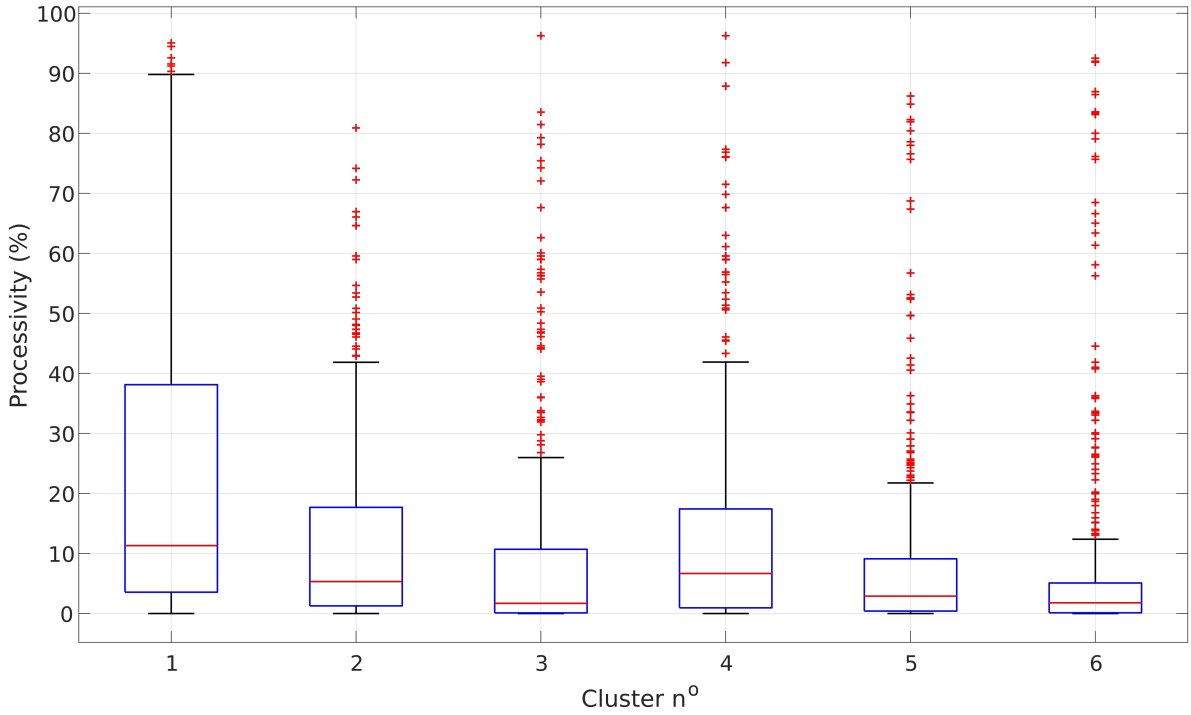


Figure 5.30: Inferred processivity of each gene in the 6 clusters determined in Figure 5.27. The processivity is defined as the proportion of polymerases which reach the end of the gene instead of undergoing early termination in the simulation sets defined by the parameter regimes which fit the mNET-seq profiles (Figures 5.28a & 5.28b).

Why transcription is such a faulty process is open to debate, but may reflect a number of "pioneer" rounds of transcription which occur at each gene before processive elongation begins. This may also appear as "bursting" transcription, with failed transcription events occurring so early into the gene that the RNA molecules transcribed would not be long enough to be detected by methods such as RNA FISH or MS2-loops. Then if successful transcription events do occur in clusters, this would result in multiple elongating polymerases at a single locus as observed in bursting studies (Lenstra *et al.* 2015, Zenklusen *et al.* 2008, Corrigan *et al.* 2016). While a slightly semantic argument, this does highlight the possibility for bursting transcription and faulty initiation to appear as the same phenomenon depending on the technique used.

5.8.3 CDK9 controls mobility of promoter-proximal polymerase

While we have been able to directly investigate the effects of removing the DSIF component Spt4 in *S. cerevisiae*, the same experiment is not possible in HeLa cells. Using the drug DRB, we are able to inhibit the phosphorylating activity of the pTEFb component

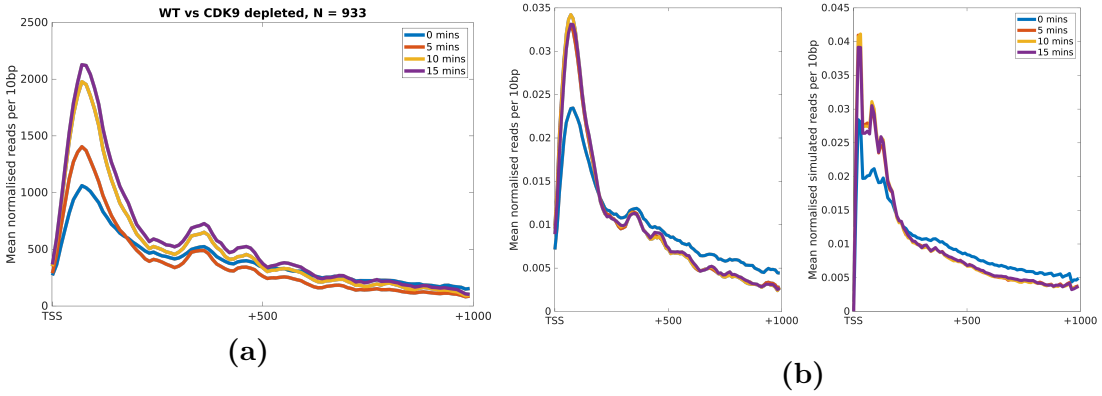


Figure 5.31: Metagenes of raw data showing distributions of transcriptionally-engaged polymerase before and after 5, 10 & 15 minutes treatment with CDK9-inhibitor DRB. Shown are the metagenes of the raw reads (left), reads normalised per gene (center) and metagenes of fitted distributions (right). Experiments performed by Michael Tellier from the lab of Shona Murphy.

CDK9 in HeLa cells. The phosphorylation of SPT5 by CDK9 marks the release of the polymerase from the promoter-proximal region, converting DSIF from a negative to a positive elongation factor and signals the release of negative elongation factor NELF from the RNA PolII complex. In this section we aim to determine the effects of inhibiting CDK9 on the dynamics of RNA PolII and whether these effects change depending on the length of treatment.

The inferred results from simulating the mNET-seq distributions confirm that the parameter shifts are conserved between 5, 10 & 15 minutes of treatment with DRB and do not become more or less pronounced over time, as is indicated by the metagene profiles in Figure 5.31. Inferred parameters from simulating mNET-seq after 5 & 10 minutes of DRB treatment are shown in Appendix 7.5.

Inhibition of the activity of CDK9 by the drug DRB results in almost immediate changes to the kinetics of RNA PolII at the 5' end of the gene. Given the annotated role of CDK9 in phosphorylating DSIF component hSPT5, resulting in proximal-pause release, it is unsurprising that the effects of inhibiting CDK9 are seen mostly at the start of the gene. We observe a reduced ratio of moving polymerase while in the first window of the gene and as a potential result of this, observe lower processivity at the majority of transcribed genes, which is constistent with a recent study showing treatment with DRB in stressed cells resulted in early termination of RNA PolII at the 5' end of genes (Erickson *et al.* 2018). This suggests a mechanism of promoter-release whereby activity of CDK9 allows polymerase to escape from the 5' region of the gene, otherwise the polymerase

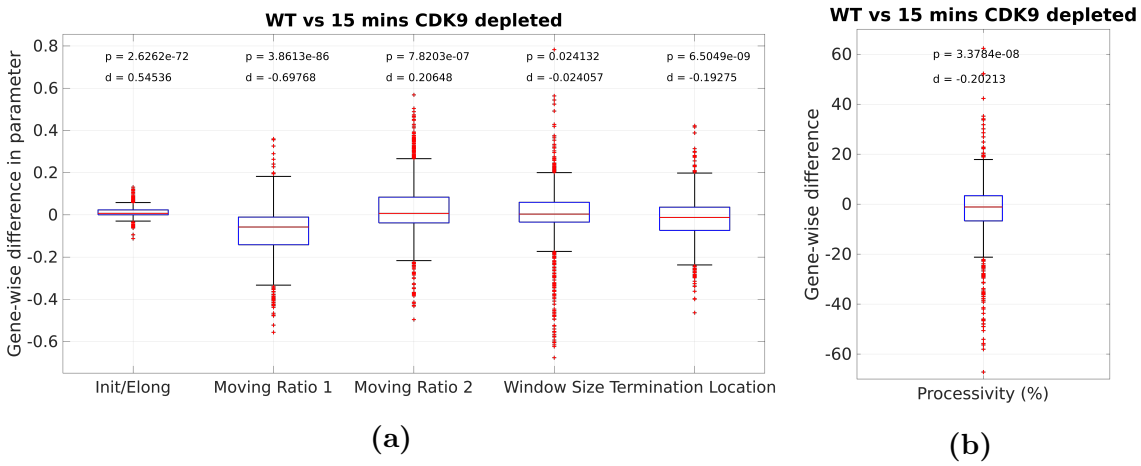


Figure 5.32: Output from simulation results comparing wild-type and DRB-treated HeLa cells. Shown are the differences in inferred parameters at each gene after 15 minutes of treatment with DRB, a CDK9-inhibitor (5.32a) and the differences in processivity of the parameters for each gene (5.32b).

undergoes early termination.

We furthermore see an increase in the ratio of initiation to elongation ratio, creating a larger steady-state number of polymerases on the gene and a 5'-shift in the location of early termination after CDK9 inhibition.

5.9 Discussion

We have shown here that the distinct profiles of polymerase seen across the *S. cerevisiae* and human genome as determined in Chapter 4 are related to distinct kinetics of transcription. Perhaps the most striking observation is that under this hypothesis-blind approach we identified a consistent model of transcription that was found to best explain the polymerase distributions of both *S. cerevisiae* and HeLa cells. This goes against previous reports that there is no pausing in *S. cerevisiae* and that the mechanisms of elongation differ between the two (Kwak & Lis 2013, Mayer *et al.* 2010, Booth *et al.* 2016), although these conclusions are often based on low-resolution techniques (ChIP-seq or GRO-seq) or invasive protocols requiring un-physiological conditions through isolation of nuclei and *in vitro* transcription assays (PRO-seq or GRO-seq).

The process of polymerase elongation from the start to the end of the gene is clearly complex, made up of the polymerase undergoing dynamic transitions from a moving state through multiple static states, creating complex distributions at each gene. Here, we

have shown that these dynamic transitions of the polymerase are regulated by a number of factors, acting to both assist and hinder the process of transcription elongation. As has been previously described, an early checkpoint in the transcription process is reflected in the model as two distinct dynamic regions, with a switch-point controlled by a number of factors (Laitem *et al.* 2015, Booth *et al.* 2018, Fitz *et al.* 2018).

Modifications directly made to RNA PolII are shown here to affect the elongation process. Removing the possibility for phosphorylation of the Tyrosine_1 residue of the CTD resulted in a much more mobile polymerase in the promoter-proximal region, highlighting a role for the phosphorylation of the CTD in regulating the more static state immediately downstream of the TSS. The modifications to the CTD rely on the recruitment of factors to phosphorylate these residues, therefore this operation is not performed in isolation, but is likely done with other elongation factors responsible for regulating the dynamics of polymerase, or recruited as a result of the CTD phosphorylation.

The elongation factors studied broadly fall into two categories, with Spt4 and Dst1 working to assist the RNA PolII complex through the gene and Paf1 acting as a pause-factor hindering polymerase progression. Dst1 was found to reduce the amount of time the polymerase spends in a stalled or backtracked state in the promoter-proximal region of the gene, similar to Spt4. Paf1, however was found to act in an opposing manner, restricting the elongation of the polymerase through this region, potentially keeping the polymerase in a stalled state downstream of the TSS in order to undergo maintenance checks. In a similarly antagonistic fashion, Spt4 was found to act to reduce the ratio of initiation to elongation rate, with Paf1 increasing it. A smaller initiation-elongation ratio can be thought to allow polymerases to clear the promoter region of the gene, and was shown to produce more even profiles under synthetic tests in Figure 5.8. The only aspect of transcription elongation towards which all three investigated elongation factors appeared to offer the same contribution was in determining the location of early termination. All three appeared in general to act to move this location further downstream, into the gene body, establishing a positive role for these factors in maintaining the contact between the polymerase and the DNA, increasing the probability of the polymerase reaching the end of the gene.

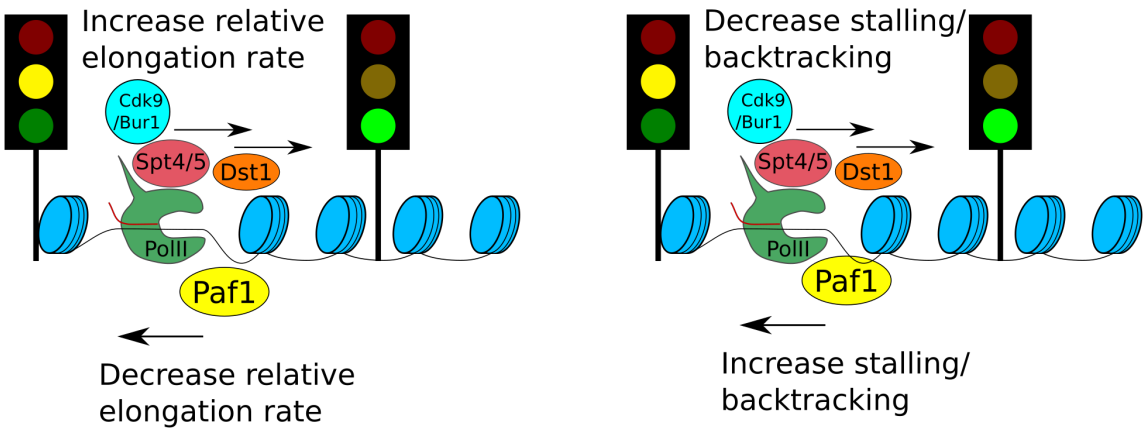


Figure 5.33: A summary of the effects on the polymerase of the studied factors associating with the RNA PolII complex. Both Spt4 and the phosphorylating activity of Cdk9 on Spt5 were found to decrease the mean time the polymerase spends in a moving state and also increases the elongation rate relative to the initiation rate. Dst1 was shown to act in a similar manner to the DSIF components, whereas Paf1 was found to increase the time the polymerase spent in either a stalled or backtracked state and reduce the elongation/initiation ratio.

The histone writers investigated all had modest effects on the dynamics of transcription, but in general acted on much wider windows than the elongation factors. The histone methyltransferases Set1 and Set2 were both found to contribute towards the average time the polymerase spent in a moving state, but were not restricted to either window, working throughout the gene. Spp1, the protein responsible for tri-methylation of H3K4, showed no consistent or large impact on transcription dynamics, suggesting H3K4me3 may have roles in the cell outside of transcriptional control, despite being associated with highly-expressed genes (Santos-Rosa *et al.* 2002, Howe *et al.* 2017, Carrozza *et al.* 2005). The subunits Eaf3 and Rco1 from the HDAC Rpd3S showed similar behaviours to each other in their control of transcription, however with Eaf3 acting mainly in the second, distal window of transcription, potentially reflecting compensatory activity of the HAT NuA4. Both subunits acted to assist the polymerase complex by reducing time spent in stalled or backtracked states. Only the H3K4 methyltransferase Set1 was found to contribute to the location of early termination, working to move it closer to the promoter region, antagonistic to the elongation factors. This is in line with similar observations made in a number of human cell lines that the elongation rates of long genes correlate with the levels of H3K79me2 and H4K20me1, but show no correlation with H3K36me3 (Veloso *et al.* 2014). Here, we have found similar results in *S. cerevisiae* through non-invasive

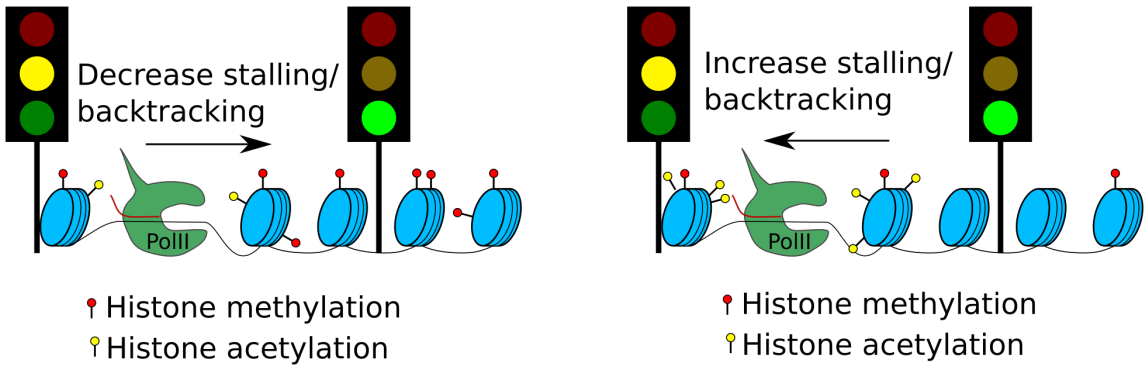


Figure 5.34: A summary of the effects of histone acetylation and methylation on polymerase activity. Methylation of H3K4 and H3K36 by Set1 and Set2 were found to assist the passage of the polymerase through the gene. Higher levels of acetylation in the promoter region, which would be removed by the de-acetylation complex Rpd3S, results in a slower polymerase through increased stalling/backtracking.

methods, not requiring any labelling by alternative NTPs. This further suggests that the observations made here would be conserved in higher eukaryotes, with transcription elongation control a conserved phenomenon involving elongation factors and the chromatin.

The deletion of the cytoplasmic exonuclease Xrn1 resulted in a more efficient elongation process, with higher processivity at the majority of genes. This is in line with observations made by the Cramer lab from cDTA analysis of the synthesis and turnover of transcripts upon deletion of Xrn1. The faster production of RNA observed in Sun *et al.* 2013 could therefore be explained through higher processivity by influencing the elongation process, rather than directly affecting the initiation of transcription events. This argues against Xrn1 as a transcription factor, affecting gene regulation in the nucleus through indirect mechanisms, most likely through imbalance of elongation factors or chromatin remodellers through improper turnover of transcripts.

One caveat of the work carried out here is that we cannot rule-out a number of indirect effects on transcription as a result of deleting these factors. Certainly in the lab it was observed that the *stp4*Δ and *xrn1*Δ strains were slow growing and do not respond well to stress. While the elongation factors are known to associate with the elongation complex, modifying the chromatin or degradation machinery may be affecting transcription through other pathways such as disrupting the levels of key transcription and elongation factors in the cell. Deleting these factors certainly affects the dynamics of transcription, however

it is beyond the scope to determine the mechanisms, whether direct or indirect, that determine these changes.

5.9.1 Transcription processivity is determined by many factors

One major conclusion from the modelling of NET-seq data is the output of transcription processivity can be estimated from the shape of the distributions of polymerase. We found that the processivity of a gene closely followed the distributions of RNA PolII across the gene, with a higher proportion of polymerase at the 5' end of the gene indicative of a lower processivity. Dst1 and Spt4 were both found to assist polymerase processivity, most likely by alleviating the RNA PolII complex from stalled or backtracked blockages to the elongation process. Paf1, however, was found to reduce processivity, with the most likely role of Paf1 being to retain the polymerase in the promoter-proximal region in a stalled or paused state, as suggested in Chen *et al.* 2015. This sets up a situation where elongation factors act antagonistically to both increase and decrease processivity at genes, potentially to determine the "sweet-spot" of overall balance of transcripts in the cell. Results are emerging to suggest that factors exist to initiate early termination of elongation near the 5' ends of genes. The chromatin remodeller Ino80 has been shown to target RNA PolII for degradation as an RNA-independent mechanism to terminate early elongation in *S. cerevisiae* (Lafon *et al.* 2015) without the need for poly-adenylation or degradation of the RNA via the NNS pathway.

Furthermore, we identified the chromatin as a determinant of polymerase processivity, with increases in acetylation and decreases in methylation resulting in lower processivity of genes. The link between methylation and acetylation through the recruitment of HDACs by histone methylation suggests that the reduction in processivity through decreased histone methylation may be through increased residual acetylation at these genes, or a direct effect through another factors recruited by the chromatin. Either way, this work outlines the importance of the chromatin and the make-up of the elongation complex in establishing the correct levels of processivity of each gene, ultimately dictating the production rate of transcripts from each gene and the amount of effort required regarding the number of initiation events per transcript produced.

Chapter 6

Conclusions and future work

6.1 Summary of results

The work presented here aims to elucidate the true underlying nature of the dynamics of gene expression and how different regulatory factors work together to establish these dynamic regimes in eukaryotic cells. The definition of gene expression has become more complex over time, moving beyond a simple of view of binary on/off switches at genes, to the response times of genes to external cues and the steps involved in fine-tuning the production and processing of RNA from a DNA template.

6.1.1 Nucleosomes are highly dynamic upon changes of gene expression

In Chapter 2, we observed that when *S. cerevisiae* cells undergo a shift in carbon source from glucose to galactose, the expression profiles are altered rapidly on a genome-wide scale. This change in expression is coupled with a change in the underlying chromatin around the promoter of each gene. At genes which undergo activation upon addition of galactose containing media we observed an opening of the NDR around the TSS, with the +1 nucleosome moving into the gene body and the opposite effect at repressed genes.

During the Yeast Metabolic Cycle, it has been shown that activation in the oxidative phase results in the coordinated expression and shifting of the +1 nucleosome at a small number of roughly 50 genes (Nocetti & Whitehouse 2016). The work presented here shows

that this phenomenon is much more widespread, and the changes seen during this narrow phase of the YMC may be more representative of a sharp change in cell metabolism, creating changes in the nucleosome distributions at key genes. Upon the rapid change of media presented in Chapter 2, the chromatin exhibits a genome-wide relaxation of the chromatin around the TSS, but the changes to the +1 nucleosome are expression-dependent and appear to begin relaxing to the previous positions after 60 minutes in galactose.

Cells are constantly under changing external conditions, being asked to metabolise different chemicals and respond to stresses or cues to enter different states. Here, we showed that such changes in conditions result in altered nucleosome arrangement which occurs within 15 minutes of the external signal. The arrangement of the nucleosomes on the DNA create varying degrees of torsional stress and bendability, which may feed back to the transcription factors, PIC and RNA PolII, which dictate the expression levels of each gene. The work in this Chapter suggests a relaxed NDR as a universal response to external stress, with the position of the +1 nucleosome a good predictor of the output of RNA at each gene across the population of cells.

6.1.2 Chromatin acts to modulate transcription dynamics

In Chapter 3 we investigated how the effects of antisense transcription and histone modifications affect the dynamics of transcription in *S. cerevisiae*. While in Chapter 2 we determined a link between the positions of nucleosomes and absolute levels of expression, we have no information about how the post-translational modifications of the nucleosomes are involved in this process.

While non-coding antisense transcription is widespread across eukaryotic organisms (Mellor *et al.* 2016), the roles of this process are fairly poorly understood. It has been shown recently that the levels of antisense transcription at genes correlate with the post-translational modifications of the histones (Murray *et al.* 2015, Brown *et al.* 2018). As the sense and antisense transcription units elongate through the same chromatin template, we may expect one to affect the other.

In this Chapter we determined that high levels of antisense transcription result in

reduced rates of transcription initiation and elongation/export, which is compensated for by a stabilisation of the mature transcripts. This results in the same steady-state levels of RNA in the absence of antisense transcription, however these results suggest that it is the production and turnover of these transcripts that are most affected by levels of antisense.

By removing the histone de-acetylase Set3, we were able to reproduce the dynamics of transcription that were observed in a strain with high levels of antisense. This suggests that the slowing of the transcription dynamics are likely to occur as a result of changes in the chromatin, with increased levels of histone acetylation and decreased levels of methylation resulting in a slower initiation and elongation/export. Furthermore, by removing Set3 in a high antisense strain we observed no further changes to the dynamics, suggesting a threshold on acetylation levels, or at least on the possible effect of acetylation on transcription dynamics, and a system in the cell where one mechanism can mediate the loss of the other. Indeed, genes with antisense transcription have been shown to exhibit smaller increases in acetylation levels following the loss of Set3 genome wide (Brown *et al.* 2018).

6.1.3 Transcription dynamics are distinct across eukaryotic genomes

In Chapters 3, 4 & 5 we observed that the dynamics of transcription cannot be thought of as one homogeneous process from initiation to termination which is conserved between organisms or even across the genome of an individual organism. Modelling RNA FISH data from 12 genes in *S. cerevisiae*, we observed a range of distributions of nuclear and cytoplasmic RNA across a population of cells. These differences in distributions were coupled by distinct parameters of transcription, yet there was a strong correlation between the initiation rate and elongation/export rate at each gene. This result was further confirmed by a more detailed bioinformatic and modelling approach to study the profiles of transcriptionally-engaged RNA PolII as determined by NETseq.

Moving away from the traditional metagene techniques used to visualise nascent transcription datasets, we were able to separate the profiles of polymerase into at least 3 groups in *S. cerevisiae* and 6 in HeLa cells. These different profiles of polymerase deter-

mine distinct parameter regimes of transcription, as determined by stochastic modelling of the transcription process which produces these profiles of polymerase. We found that increased ratios of the proximal-to-distal signal were indicative of increased stalling or backtracking in the 5' regions of genes and of a slower elongation rate relative to the initiation rate.

Furthermore we showed that distributions of polymerase, and particularly the degree of accumulation at the 5' ends of genes, were independent of the overall levels of expression and are consistent upon natural fluctuations in expression through the Yeast Metabolic Cycle. This indicates that the dynamics of each gene are likely to be inherent to the gene, rather than being a consequence of external cues. It is likely that this is due partly to the chromatin arrangement at each gene, but also the factors that bind to the DNA and RNA PolII complex. It has been shown that elongation factors preferentially bind subsets of genes over others, most likely determining the downstream processing of the RNA (Fischl *et al.* 2017). We would suggest further that the dynamics of the polymerase are also specific to subsets of genes, which may be a result of or lead to the alternative processing of the transcript.

Constructing multiple possible stochastic models of RNA PolII transcription, we were able to establish a best-fit model which captured the NETseq profiles in *S. cerevisiae* and mNETseq profiles in HeLa cells. Using this model of transcription, we were able to further probe the ranges of dynamics that exist across eukaryotic organisms and whether the dynamics are conserved between the simple and higher eukaryotes studied.

One immediate consequence of these findings was that these genes with more 5' steady state RNA PolII generally were much less processive than those genes with flatter, more even distributions of polymerase. This points to an increased number of early termination events that occur in these genes, with the promoter-proximal region clearly being a key area which is highly regulated. A high elongation/initiation ratio may allow the RNA PolII complex to rapidly exit this area, allowing further rounds of transcription to occur at highly processive genes, with the opposite acting to restrict future initiation events (Shao & Zeitlinger 2017) or cause large numbers of collisions, resulting in removal of polymerases causing blockages (Svejstrup 2007).

6.1.4 Many factors work cooperatively and antagonistically to determine the parameters of transcription

To further probe the roles that different factors have on the rates of the key steps of transcription, we analysed RNA FISH and NETseq data from *S. cerevisiae* strains after deletion of proteins or in HeLa cells following inhibition of CDK9 by drug treatment. Firstly, the elongation factors analysed demonstrated that not all factors act in the same manner, and may not even necessarily assist the RNA PolII complex through the gene. We found that the DSIF component Spt4 and the backtrack release factor Dst1 (TFIIS) both acted to assist the polymerase through the promoter-proximal region of the gene by increasing the ratio of time the polymerase spends in a moving state as opposed to stalled or backtracked. As both factors are associated with pause and backtrack release (Wada *et al.* 1998a, Mason & Struhl 2005) this may not be surprising. The outcome of this result is that both Spt4 and Dst1 work to increase polymerase processivity and the efficiency of the elongation process to produce a fully-formed transcript.

Contrary to this, we observed that Paf1 works to maintain the polymerase in a stalled state in the promoter-proximal region of the gene. This suggest that Paf1 acts antagonistically to DSIF in *S. cerevisiae*, working to maintain any promoter-proximal pauses, potentially to ensure the elongation complex has time to accumulate the correct factors, modifications and proceed effectively into processive elongation, in line with observations from the Shilatifard lab (Chen *et al.* 2015).

We also found that the RNA exonuclease Xrn1 inhibits the elongation of a large number of genes in *S. cerevisiae*, with a higher processivity of genes seen in *xrn1*Δ strains. This is in line with observations made by the Cramer lab, with Xrn1 affecting both the production and turnover rate of transcripts (Sun *et al.* 2013). We are not able to distinguish the exact steps Xrn1 takes in modulating elongation dynamics, but would speculate that they are likely to occur through an imbalance in factors created as a result of disrupting the RNA degradation pathway in the cells.

6.1.5 Histone modifications modulate polymerase elongation

In Chapters 3, 4 & 5 we aimed to further elucidate the roles of histone modifications in governing transcription.

Acetylation works to inhibit elongation

In Chapter 3 we determined that increased histone acetylation, whether brought about by levels of antisense transcription or the loss of a de-acetylase, resulted in slowed dynamics of transcription. This was further confirmed by results in Chapter 5, where modelling NETseq data following loss of factors from the histone de-acetylation complex Rpd3S resulted in a less mobile polymerase. This suggests that there is interplay between the chromatin and the RNA PolII Complex, whether direct or indirect, where maintaining low levels of histone acetylation are conducive to more processive transcription.

However, allowing high levels of acetylation may allow for further factors to bind both the DNA and the elongation complex, for example the HDACs, which may have further downstream consequences on further rounds of transcription (Lee *et al.* 2018). This may be of particular importance in systems where cells must undergo activation and repression of genes, such as the Yeast Metabolic Cycle. Furthermore, more concise expression of genes may result in more homogeneous expression across a population of cells, reducing the levels of noise across a population (Huber *et al.* 2016, Steensels & Verstrepen 2016).

However, the role of histone acetylation, and particularly of HDACs may lie outside of regulating the closest transcription unit in *cis*. As has been suggested, the roles of the modifications to the chromatin via Set2 and Rpd3S may be to regulate the transcription of non-coding transcripts originating from novel sources, rather than regulate sense transcription in any way as previously reported (Churchman & Weissman 2011, Kim *et al.* 2016). Given the crowded nature of the *S. cerevisiae* genome, it is difficult to take genes in isolation, with many shared chromatin templates and 3-dimensional structures not studied here.

Methylation, but not all methylation, encourages polymerase elongation

In Chapter 3, we revealed that loss of Spp1, and H3K4me3, did not affect the dynamics of transcription initiation or elongation at 12 genes when analysing RNA FISH data. This result was confirmed after modelling NETseq data from cells without Spp1 on a genome-wide scale, suggesting a consistent finding that H3K4me3 is not instructive for transcription or the dynamics of the polymerase (Howe *et al.* 2017).

We did however observe an effect on the dynamics of elongation after the deletion of the histone methyltransferases Set1 and Set2. As Set1 methylates H3K4 and Set2 H3K36 we may not necessarily expect the two modifications to have the same effect as each other, but we observed increased stalling of the polymerase upon loss of methylation of these histone residues. This points to the mono- and di-methylation of H3K4 acting to assist the polymerase through the gene body, either directly or through the recruitment of other factors and similarly with the methylation of H3K36.

Modifying the chromatin at genes through methylation, acetylation and other post-translational modifications creates an ever-changing environment through which the RNA PolII complex must transcribe. Together, they point to an antagonistic network of interactions, yet do not all operate at the same time. The placement of the methylation marks on H3K4 by the act of transcription may lead to further methylation through repeated rounds of transcription (Soares *et al.* 2017, Lee *et al.* 2018). Similarly the recruitment of histone de-acetylation complexes by histone methylation (Carrozza *et al.* 2005, Venkatesh *et al.* 2012) may further alleviate polymerase stalling by reducing the levels of histone acetylation on the chromatin (Ginsburg *et al.* 2014).

6.1.6 RNA PolII transcribes in the face of multiple obstacles

The main aim of transcription by RNA PolII is to produce transcripts to then be translated into functional proteins. The numbers of transcripts corresponding to each gene ultimately dictates the balance of proteins in the cell, which must be tightly regulated in order for the cells to react to changing external conditions. Here we have shown that the dynamics of the polymerase at each gene are distinct and the procession of the polymerase is both helped and hindered by interacting proteins and the surrounding chromatin and

transcriptional environment.

Genes may have to be transcribed in spite of, rather than as a result of, the other processes occurring around the genic template on the DNA. The degree of stalling and backtracking at the beginning of genes hinders the processivity of the polymerase, yet is independent of the amount of RNA produced at each gene. This suggests that cells not only have genes with high and low expression, but also genes with high and low degrees of stalling and the two are not related. Therefore the degree of activation of each gene is not only a function of the required amount of RNA, but the inherent efficiency of each gene. A highly processive gene only requiring a small number of steady-state transcripts per cell may only exhibit rare initiation events, whereas a gene with lots of RNA but a low processivity will require a disproportionately high number of initiation events over a fixed time, not all of which will produce an RNA molecule.

The balance of histone acetylation, methylation, the levels of antisense transcription at a gene and the different factors that bind to the elongation complex all work together to create these dynamics of transcription inherent to each gene. The work presented here suggests that each of these factors do not have the role of activating or repressing gene expression, but rather work to fine-tune the activity of the polymerase complex. Under different conditions this may allow the cell to modulate the levels of histone modifications on a genome-wide scale to simply shift the degrees of stalling or abortive transcription to change the overall distributions of RNA, without the need to drastically redistribute the polymerase in the cell, which may be more costly.

6.2 Future work

The work presented here sets up the possibility to investigate a number of key steps in the transcription process. As more sophisticated experimental techniques are developed, these should also be coupled with in-depth numerical analysis through bioinformatic and modelling techniques.

The further development of mNETseq by the Proudfoot lab, pulling down not only RNA PolII, but each phosphorylation modification of the CTD adds more dimensions to the data (Nojima *et al.* 2015, Schlackow *et al.* 2017). This gives scope to extend the models

presented to also incorporate the kinetic transitions made by the polymerase in terms of its modifications and couple this with any spatial transitions made by the elongation complex, particularly when studying the move from early to processive elongation.

These observations should also be combined with the modifications to the elongation complex regarding the proteins known to associate and dissociate with the polymerase throughout the many stages of transcription. Taking TEFseq and PAR-ClIP data which shows where certain factors bind the polymerase and nascent RNA will also pin-point the dynamic consequences of these factors in the elongation process (Fischl *et al.* 2017, Battaglia *et al.* 2017).

Not presented in depth here is the difference in polymerase profiles exhibited at genes when determined by NETseq and PROsesq. As the two techniques in theory capture the same phenomenon, it is not immediately obvious why NETseq profiles generally show a higher proportion of reads at the start of genes, whereas PROseq show even distributions throughout the gene body. The main difference between the two techniques is the addition of the run-on reaction in PROseq, perhaps suggestive of a different form of polymerase at the 5' ends of genes, which cannot elongate under the *in vitro* elongation conditions. That the same is not true in human cells also indicates a divergence between dynamic states of the polymerases of the two organisms, but requires more in-depth analysis.

Finally, the work presented here has mainly focused on the 5' ends of genes, looking at the roles of factors in establishing the dynamics of initiation, promoter release and processive elongation. The final stage in the transcription, termination, is also a highly regulated step in the production of RNA, which is also speculated to undergo dynamic transitions involving pausing around the termination site (Tellier *et al.* 2016, Fischl *et al.* 2017, Booth *et al.* 2018). Building similar models focused on the 3' ends of genes would give further scope to study the ranges of dynamics in this region and the roles of individual factors on the key steps in the final step in transcription.

Chapter 7

Appendix

7.1 Nucleosome locations of YMC classes of genes upon galactose shift

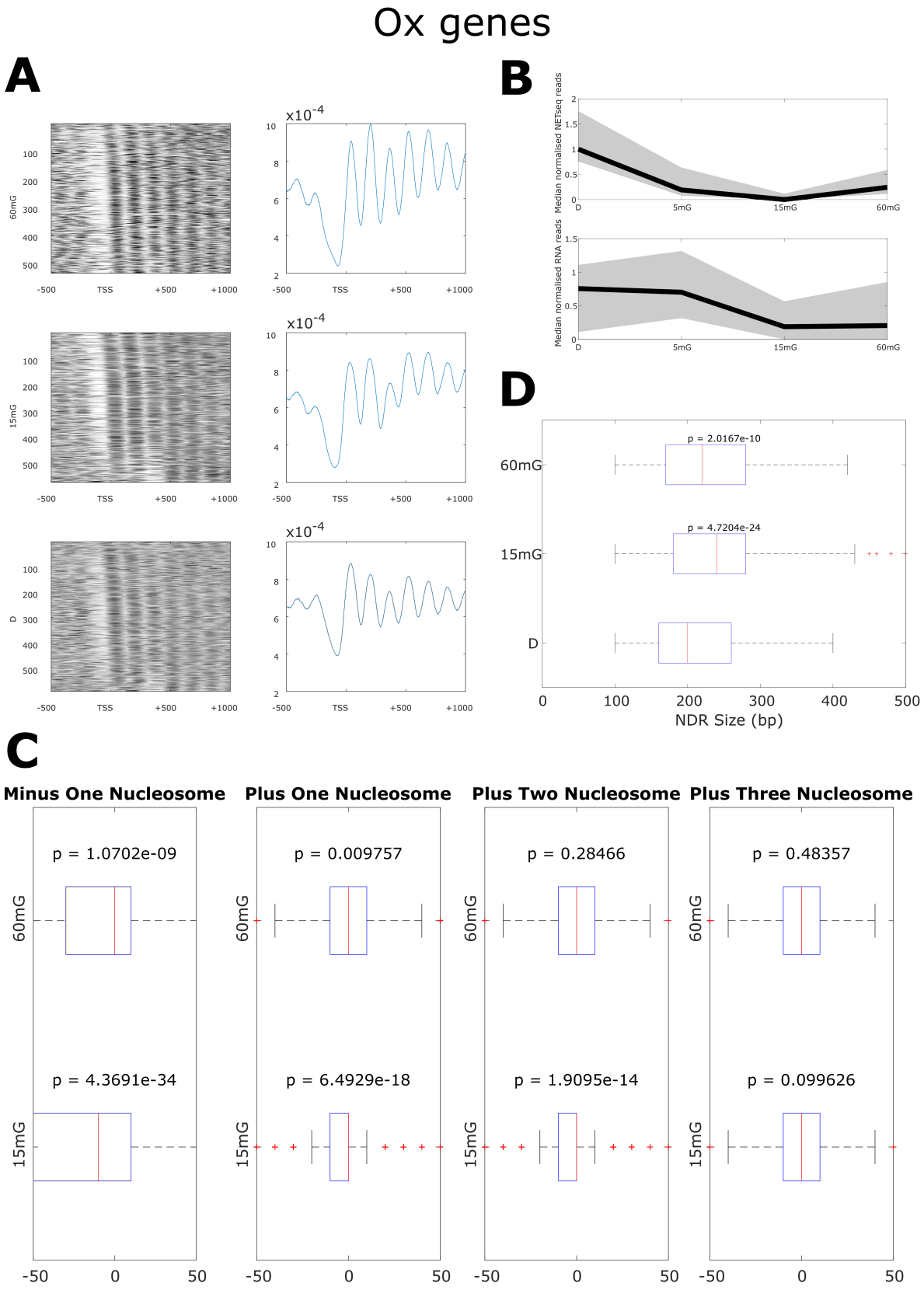


Figure 7.1: Positions of nucleosomes at the Ox genes under shift in media from glucose to galacose. Shown (**A**) are the normalised MNase profiles of each gene in the Ox cluster at the three time-points and the averaged profile across all genes in the Ox cluster. (**B**)The location of the -1, +1, +2 and +3 nucleosome after 15 and 60 minutes in galactose relative to the glucose location. (**C**) The size of the nucleosome depleted region (NDR), defined as the distance between the -1 and +1 nucleosome centres in the three time-points.

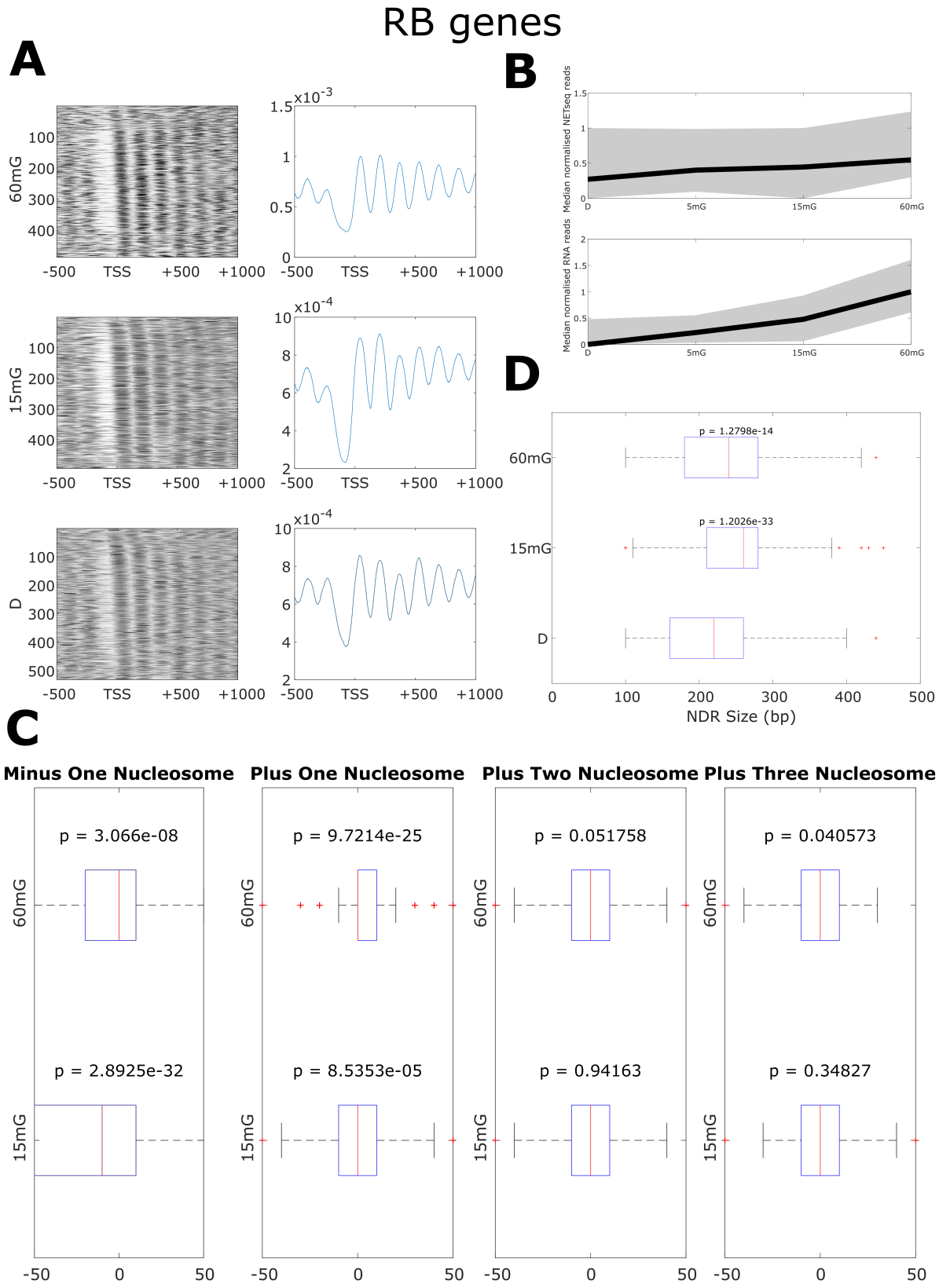


Figure 7.2: Positions of nucleosomes at the RB genes under shift in media from glucose to galacose. Shown (**A**) are the normalised MNase profiles of each gene in the RB cluster at the three time-points and the averaged profile across all genes in the RB cluster. (**B**)The location of the -1, +1, +2 and +3 nucleosome after 15 and 60 minutes in galactose relative to the glucose location. (**C**) The size of the nucleosome depleted region (NDR), defined as the distance between the -1 and +1 nucleosome centres in the three time-points.

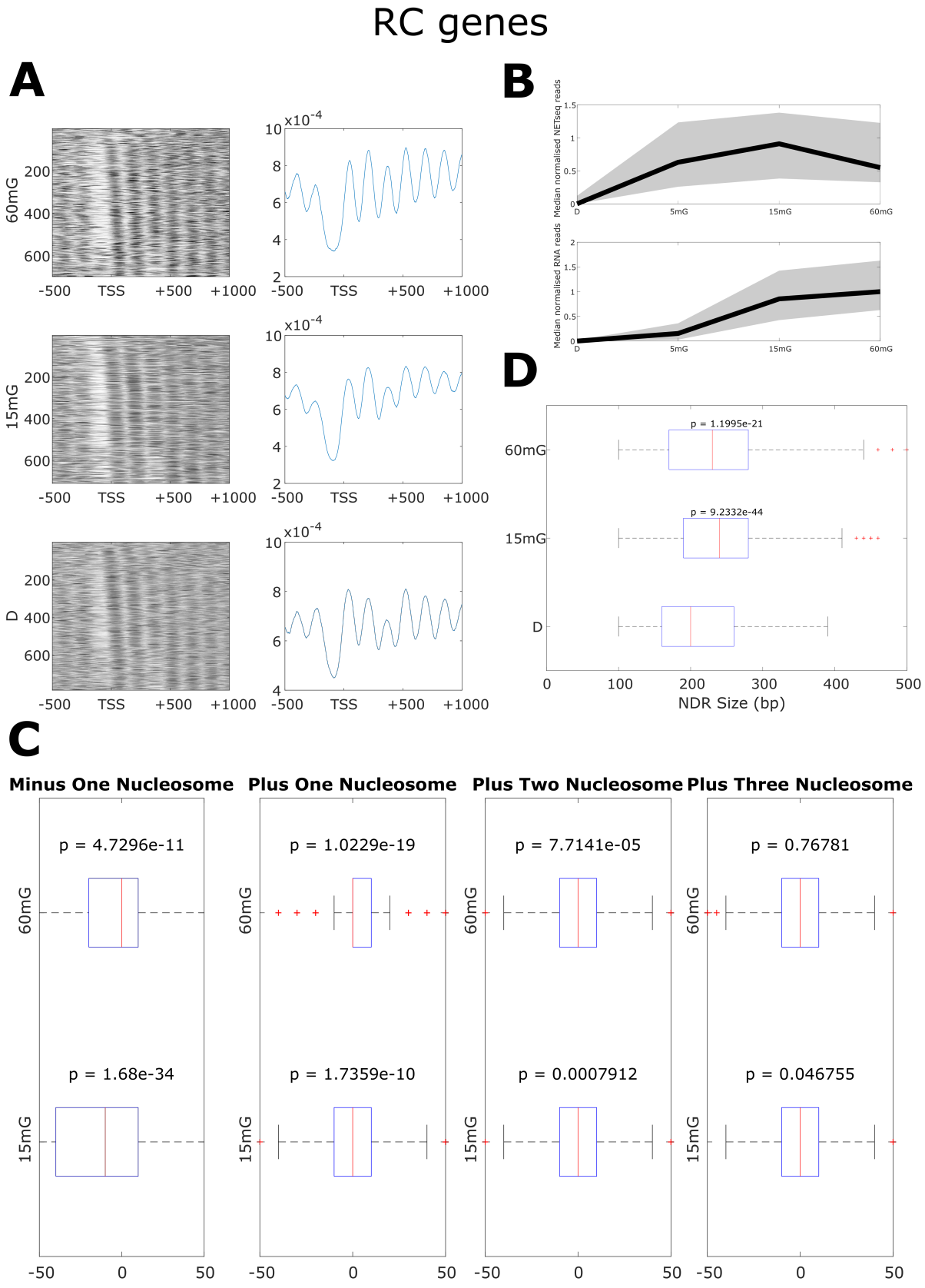


Figure 7.3: Positions of nucleosomes at the RC genes under shift in media from glucose to galacose. Shown (**A**) are the normalised MNase profiles of each gene in the RC cluster at the three time-points and the averaged profile across all genes in the RC cluster. (**B**)The location of the -1, +1, +2 and +3 nucleosome after 15 and 60 minutes in galactose relative to the glucose location. (**C**) The size of the nucleosome depleted region (NDR), defined as the distance between the -1 and +1 nucleosome centres in the three time-points.

7.2 Derivation of the probability generating function for bursting protein production

The following is taken from Peccoud & Ycart 1995 with a number of intermediary lines added for clarity and notation made consistent with Chapter 3.

For the system:

$$Inactive\ Promoter \underset{\beta}{\overset{\alpha}{\rightleftharpoons}} Active\ Promoter \tag{7.1}$$

$$Active\ Promoter \xrightarrow{\gamma} Active\ Promoter\ +\ Protein \tag{7.2}$$

$$Protein \xrightarrow{\delta} \emptyset \tag{7.3}$$

the transition probabilities are given by:

$$p(a,n,t) = \mathbb{P}\{Promoter = a \in [0,1], Protein = n \in [0,\infty), time = t\} \tag{7.4}$$

$$p_0(n,t) = \mathbb{P}\{Protein = n \in [0,\infty), time = t | Promoter = 0\} \tag{7.5}$$

$$p_1(n,t) = \mathbb{P}\{Protein = n \in [0,\infty), time = t | Promoter = 1\} \tag{7.6}$$

where the promoter is modelled as being in an active ($a = 1$) or inactive ($a = 0$) state. We separate the transitions into those with active and inactive promoters as follows:

$$\frac{dp_0}{dt}(n,t) = \beta p_1(n,t) + \delta(n+1)p_0(n+1,t) - (\delta n + \alpha)p_0(n,t) \tag{7.7}$$

$$\frac{dp_1}{dt}(n,t) = \alpha p_0(n,t) + \delta(n+1)p_1(n+1,t) + \gamma p_1(n-1,t) - (\delta n + \gamma + \beta)p_1(n,t) \tag{7.8}$$

where $p_1(-1,t) = 0$ for all values of t. By defining $\phi_0,\ \phi_1$ to be the long-run distribution of protein, in the limit $t \to \infty$ we obtain:

$$(7.7) \implies 0 = \beta\phi_1(n) + \delta(n+1)\phi_0(n+1) - (\delta n + \alpha)\phi_0(n) \tag{7.9}$$

$$(7.8) \implies 0 = \alpha\phi_0(n) + \delta(n+1)\phi_1(n+1) + \gamma\phi_1(n-1) - (\delta n + \gamma + \beta)\phi_1(n) \tag{7.10}$$

Defining the stationary distribution probability generating functions:

$$g_0(z) = \sum_{n=0}^{\infty} z^n \phi_0(n) \tag{7.11}$$

$$g_1(z) = \sum_{n=0}^{\infty} z^n \phi_1(n) \tag{7.12}$$

we obtain a set of differential equations for the probability generating functions by multiplying equations by z^n and summing over possible values of n:

$$(7.9) \implies 0 = \sum_{k=0}^{\infty} \beta z^n \phi_1(n) + \sum_{k=0}^{\infty} \delta(n+1) z^n \phi_0(n+1) - \sum_{k=0}^{\infty} (\delta n + \alpha) z^n \phi_0(n) \tag{7.13}$$

$$\implies 0 = \beta g_1(z) + \delta g_0'(z) - \delta z g_0'(z) - \alpha g_0(z) \tag{7.14}$$

$$\implies \delta(z-1) g_0'(z) = \beta g_1(z) - \alpha g_0(z) \tag{7.15}$$

$$(7.10) \implies 0 = \sum_{k=0}^{\infty} \alpha z^n \phi_0(n) + \sum_{k=0}^{\infty} \delta(n+1) z^n \phi_1(n+1) + \sum_{k=0}^{\infty} \gamma z^n \phi_1(n-1) - \sum_{k=0}^{\infty} (\delta n + \gamma + \beta) z^n \phi_1(n) \tag{7.16}$$

$$\implies 0 = \alpha g_0(z) + \delta g_1'(z) + \gamma z g_1(z) - \delta z g_1'(z) - (\gamma + \alpha) g_1(z) \tag{7.17}$$

$$\implies \delta(z-1) g_1'(z) = \alpha g_0(z) + \gamma(z-1) g_1(z) - \alpha g_1(z) \tag{7.18}$$

Furthermore, we have that $g_0(1)$ and $g_1(1)$ are the long-run probability of the promoter being inactive or active, respectively. As such, these values are given by:

$$g_0(1) = \frac{\beta}{\alpha + \beta} \tag{7.19}$$

$$g_1(1) = \frac{\alpha}{\alpha + \beta} \tag{7.20}$$

Differentiating (7.15) and substituting $g_1'(z)$ from (7.18) and $g_1(z)$ from (7.15) allows for

the formation of second order ordinary differential equation for $g_0(z)$:

$$(7.15) \implies \quad \delta(z-1)g_0''(z) + \delta g_0'(z) = \beta g_1'(z) - \alpha g_0'(z) \tag{7.21}$$

$$(7.15) \implies \quad g_1(z) = \frac{1}{\beta}\left(\delta(z-1)g_0'(z) + \alpha g_0(z)\right) \tag{7.22}$$

$$\implies \quad g_1'(z) = \frac{1}{\beta}\left(\delta g_0'(z) + \delta(z-1)g_0''(z) + \alpha g_0'(z)\right) \tag{7.23}$$

$$\implies \quad \delta^2(1-z)g_0''(z) - \delta(\alpha+\beta+\delta+\gamma(1-z))g_0'(z) + \alpha\gamma g_0(z) = 0 \tag{7.24}$$

Setting $x = 1 - z$ and $h_0(x) = g_0(1-z)$ gives:

$$\delta^2 x h_0''(x) - \delta(\alpha+\beta+\delta+\gamma(1-z))h_0'(x) + \alpha\gamma h_0(x) = 0 \tag{7.25}$$

The values of the derivatives evaluated at zero are given by:

$$h^{(n)}(0) = (-1)^n \frac{c(c+b)...(c+b(n-1))}{a(a+1)...(a+n-1)} \frac{\beta}{\alpha+\beta},\ n \geq 1 \tag{7.26}$$

where

$$a = \frac{\alpha+\beta+\delta}{\delta}, \quad b = \frac{\gamma}{\delta}, \quad c = \frac{\alpha\gamma}{\delta^2} \tag{7.27}$$

which can be shown by induction. $h_0(0) = \frac{\beta}{\alpha+\beta}$, so the probability generating function is given by:

$$g_0(z) = \frac{\beta}{\alpha+\beta}\left(1 + \sum_{n=1}^{\infty} \frac{1}{n!} \frac{c(c+b)...(c+b(n-1)}{a(a+1)...(a+n-1)}(z-1)^n\right) \tag{7.28}$$

And using (7.15), the probability generating function $g_1(z)$ is given by:

$$g_1(z) = \frac{1}{\alpha+\beta}\left(\alpha + \sum_{n=1}^{\infty} \frac{\alpha+n\delta}{n!} \frac{c(c+b)...(c+b(n-1)}{a(a+1)...(a+n-1)}(z-1)^n\right) \tag{7.29}$$

Then the probability generating function of the asymptotic distribution $\phi(n)$ is given by

$g(z) = g_0(z) + g_1(z)$:

$$g(z) = 1 + \sum_{n=1}^{\infty} \frac{\alpha + \beta + n\delta}{\alpha + \beta} \frac{1}{n!} \frac{c(c+b)...(c+b(n-1)}{a(a+1)...(a+n-1)} (z-1)^n \tag{7.30}$$

$$= 1 + \sum_{n=1}^{\infty} \frac{1}{n!} \frac{c(c+b)...(c+b(n-1)}{(a-1)a(a+1)...(a+n-2)} (z-1)^n \tag{7.31}$$

$$\text{as } \frac{\alpha + \beta + \mathrm{n}\delta}{\delta} = \mathrm{a} + \mathrm{n} - 1, \text{ and } \frac{\alpha + \beta}{\delta} = \mathrm{a} - 1$$

$$= {}_1F_1\left(\frac{\alpha}{\delta}, \frac{\alpha + \beta}{\delta}, \frac{\gamma}{\delta}(z-1)\right) \tag{7.32}$$

Where the confluent geometric function of the first kind is defined as:

$${}_1F_1(a, b, c) = 1 + \sum_{n=1}^{\infty} \frac{1}{n!} \frac{a(a+1)...(a+n-1)}{b(b+1)...(b+n-1)} c^n \tag{7.33}$$

The mean of the stationary distribution, $\phi(n)$ can thus be found by $\mathbb{E}[\phi(n)] = g'(1)$:

$$g'(z) = \sum_{n=1}^{\infty} \frac{1}{(n-1)!} \frac{c(c+b)...(c+b(n-1)}{(a-1)a(a+1)...(a+n-2)} (z-1)^{n-1} \tag{7.34}$$

$$g'(1) = \frac{c}{a-1} \tag{7.35}$$

$$= \frac{\frac{\alpha\gamma}{\delta^2}}{\frac{\alpha+\beta}{\delta}} \tag{7.36}$$

$$\text{So} : \mathbb{E}[\phi(\mathrm{n})] = \frac{\alpha}{\alpha + \beta} \frac{\gamma}{\delta} \tag{7.37}$$

7.3 Algorithm for simulation of RNA FISH data

Input: Promoter activation rate, α
Promoter inactivation rate, β
Initiation rate, γ
Elongation rate, k
Degradation rate, δ
Gene length, N

Output: Number of nuclear and cytoplasmic RNA in 10,000 cells

for cell = 1:10,000 **do**
  $t = 0$
  $Promoter = 0$
  $RNA = 0$
  $Termination\ Times = [\ ]$
  while $t < 500$ **do**
    $\phi_0 = \alpha * (1 - Promoter)$
    $\phi_1 = \beta * Promoter$
    $\phi_2 = \gamma * Promoter$
    $\phi_3 = \delta * RNA$
    $\phi = \sum_{i=0}^{3} \phi_i$
    $r_1 = randExp(\phi)$
    $t = t + r_1$
    if $t > Termination\ Times[0]$ **then**
      $t = Termination\ Times[0]$
      $RNA = RNA + 1$
      Remove first element of Termination times
    else
      $r_2 = rand(0, 1)$
      if $r_2 < \frac{\phi_0}{\phi}$ **then**
        $Promoter = 1$;
      else if $r_2 < \frac{\phi_0+\phi_1}{\phi}$ **then**
        $Promoter = 0$;
      else if $r_2 < \frac{\phi_0+\phi_1+\phi_2}{\phi}$ **then**
        $r_3 = randGamma(N, k)$
        Add termination time $t + r_3$
      else
        $RNA = RNA - 1$

```
            end if
        end if
    end while
    Nuclear RNA = size(Termination Times)
    Cytoplasmic RNA = RNA
end for
```

7.4 Algorithm for simulation of NET-seq distributions

Input: Promoter activation rate,
Promoter inactivation rate,
Initiation rate,
Mean location of TSS,
Standard deviation of TSS location,
Elongation rate,
Stalling rate (1 & 2),
Stalling release rate (1 & 2),
Backtrack rate (1 & 2),
Backtrack release rate (1 & 2),
Early termination centre,
Size of first window,
Timestep size, Δt,
PolII width,
Gene length

Output: Locations of moving, stalled and backtracked polymerases on a gene in 10,000 cells

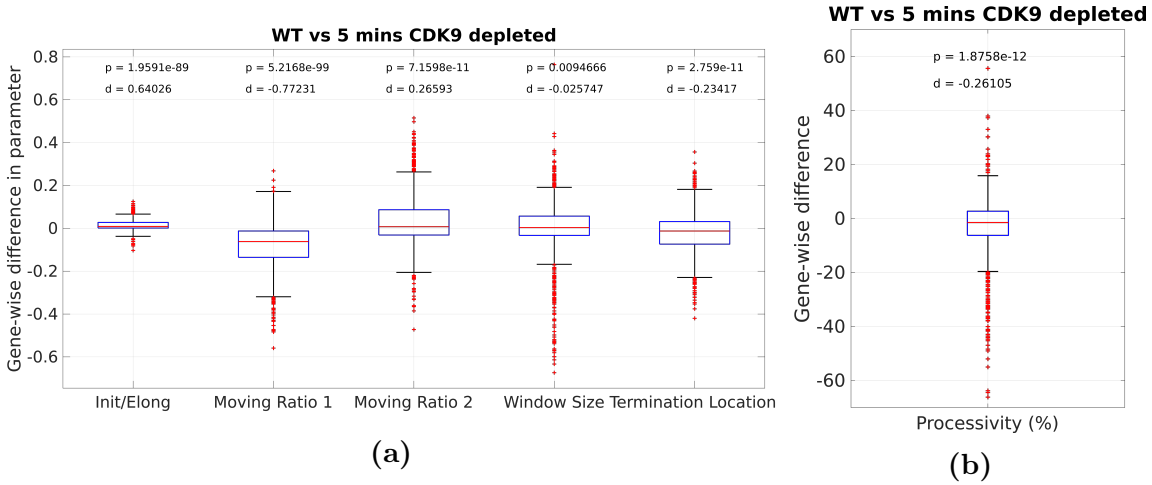
for cell = 1:10,000 **do**
  $t = 0$
  $Promoter = 1$
  $Polymerase\ positions = [\ ]$
  $Polymerase\ states = [\ ]$
  if $rand(0,1) < Initiation\ rate \times \Delta t$ **then**
    Add polymerase at position 1
    Add polymerase with state 0
  end if
  while $t < 50$ **do**
    for Number of polymerases **do**
      if $rand(0,1) < Termination\ rate \times \Delta t$ **then**
        Remove polymerase from vectors
      else
        if $Polymerase\ position < Window\ size$ **then**
          if $Polymerase\ state = 0$ **then**

```
                    if rand(0,1) < Stall_1 × Δt then
                        Polymerase state = 1
                    else
                        Advance Polymerase position by randPoi(elongation × Δt)
                    end if
                else if Polymerase state = 1 then
                    if rand(0,1) < Backtrack_1 × Δt then
                        Polymerase state = 2
                    else if rand(0,1) < StallRestart_1 × Δt then
                        Polymerase state = 0
                        Advance Polymerase position by randPoi(elongation × Δt)
                    end if
                else
                    if rand(0,1) < BacktrackRelease_1 × Δt then
                        Polymerase state = 0
                        Advance Polymerase position by randPoi(elongation × Δt)
                    end if
                end if
            else
                if Polymerase state = 0 then
                    if rand(0,1) < Stall_2 × Δt then
                        Polymerase state = 1
                    else
                        Advance Polymerase position by randPoi(elongation × Δt)
                    end if
                else if Polymerase state = 1 then
                    if rand(0,1) < Backtrack_2 × Δt then
                        Polymerase state = 2
                    else if rand(0,1) < StallRestart_2 × Δt then
                        Polymerase state = 0
                        Advance Polymerase position by randPoi(elongation × Δt)
                    end if
                else
                    if rand(0,1) < BacktrackRelease_2 × Δt then
                        Polymerase state = 0
                        Advance Polymerase position by randPoi(elongation × Δt)
                    end if
                end if
            end if
        end if
    end for
    Check overlapping polymerases
    if Collision with downstream polymerase then
        if Downstream Polymerase state > 0 then
            Upstream Polymerase state = 1
            Upstream Polymerase position =Downstream Polymerase position − PolII width
            Remove downstream polymerase
```

```
            else
                Upstream Polymerase state = 1
                Upstream Polymerase position =Downstream Polymerase position − PolII width
            end if
        else if Polymerase location > Gene length then
            Remove polymerase
        end if
        Promoter activity
        if Promoter = 1 then
            if rand(0, 1) < Promoter inact × Δt then
                Promoter = 0;
            else
                if First Promoter position > PolII width then
                    if rand(0, 1) < Initiation rate × Δt then
                        Add polymerase at position 1
                        Add polymerase with state 0
                    end if
                end if
            end if
        else
            if rand(0, 1) < Promoter act × Δt then
                Promoter = 1
                if First Promoter position > PolII width then
                    if rand(0, 1) < Initiation rate × Δt then
                        Add polymerase at position 1
                        Add polymerase with state 0
                    end if
                end if
            end if
        end if
    end while
end for
```

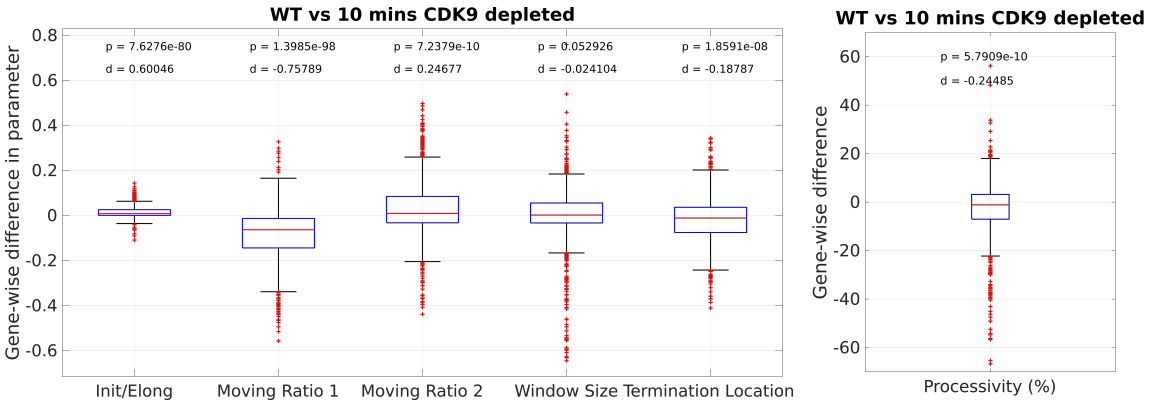
7.5 Results from simulations of mNET-seq data after treatment with DRB

The following figures show the inferred parameters from modelling mNET-seq data after 5 minutes treatment with CDK9-inhibitor DRB (Figure 7.4) and 10 minutes (Figure 7.5). The figures are presented as a comparitor to Figure 5.32 in Chapter 5, which shows the changes in parameters after 15 minutes of DRB treatment.



(a) (b)

Figure 7.4: Output from simulation results comparing wild-type and DRB-treated HeLa cells. Shown are the differences in inferred paramters at each gene after 5 minutes of treatment with DRB, a CDK9-inhibitor (7.4a) and the differences in processivity of the parameters for each gene (7.4b).



(a) (b)

Figure 7.5: Output from simulation results comparing wild-type and DRB-treated HeLa cells. Shown are the differences in inferred paramters at each gene after 10 minutes of treatment with DRB, a CDK9-inhibitor (7.5a) and the differences in processivity of the parameters for each gene (7.5b).

Publications

Brown, T. *et al.* SimBac: simulation of whole bacterial genomes with homologous recombination. *Microbial Genomics* **2,** 1–6. ISSN: 2057-5858 (2016)

Brown, T. *et al.* Antisense transcription-dependent chromatin signature modulates sense transcript dynamics. *Molecular Systems Biology* **14,** 1–21 (2018)

Bibliography

1. Adam, C. *et al.* The PHD finger protein Spp1 has distinct functions in the Set1 and the meiotic DSB formation complexes. *PLoS Genetics* **14,** 1–22. ISSN: 15537404 (2018).
2. Adelman, K. & Henriques, T. Transcriptional speed bumps revealed in high resolution. *Nature.* ISSN: 0028-0836 (2018).
3. Aguilera, A. Cotranscriptional mRNP assembly: From the DNA to the nuclear pore. *Current Opinion in Cell Biology* **17,** 242–250. ISSN: 09550674 (2005).
4. Akhtar, M. S. *et al.* TFIIH Kinase Places Bivalent Marks on the Carboxy-Terminal Domain of RNA Polymerase II. *Molecular Cell* **34,** 387–393. ISSN: 10972765 (2009).
5. Allen, B. L. & Taatjes, D. J. The Mediator complex: A central integrator of transcription. *Nature Reviews Molecular Cell Biology* **16,** 155–166. ISSN: 14710080 (2015).
6. Anderson, M. S. *et al.* Projection of an immunological self shadow within the thymus by the aire protein. *Science* **298,** 1395–1401. ISSN: 00368075 (2002).
7. Aslanzadeh, V., Huang, Y., Sanguinetti, G. & Beggs, J. D. Transcription rate strongly affects splicing fidelity and cotranscriptionality in budding yeast. *Genome Res.* **28,** 203–213. ISSN: 1088-9051 (2018).
8. Azofeifa, J. G. & Dowell, R. D. A generative model for the behavior of RNA polymerase. *Bioinformatics* **33,** 227–234. ISSN: 14602059 (2017).
9. Baejen, C. *et al.* Genome-wide Analysis of RNA Polymerase II Termination at Protein-Coding Genes. *Mol. Cell* **66,** 38–49.e6. ISSN: 10974164 (2017).
10. Barrio, M., Burrage, K., Leier, A. & Tian, H. Oscillatory regulation of hes1: Discrete stochastic delay modelling and simulation. *PLoS Computational Biology* **2,** 1017–1030. ISSN: 1553734X (2006).
11. Bartkowiak, B. *et al.* CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. *Genes & Development* **24,** 2303–2316. ISSN: 0890-9369 (2010).
12. Battaglia, S. *et al.* RNA-dependent chromatin association of transcription elongation factors and pol II CTD kinases. *Elife* **6.** ISSN: 2050084X (2017).
13. Belotserkovskaya, R. *et al.* FACT facilitates transription-dependent nucleosomes alteration. *Science* **301,** 1090–1093 (2003).
14. Birrell, G. W. *et al.* Transcriptional response of Saccharomyces cerevisiae to DNA-damaging agents does not identify the genes that protect against these agents. *Proceedings of the National Academy of Sciences* **99,** 8778–8783. ISSN: 0027-8424 (2002).
15. Boija, A. *et al.* CBP Regulates Recruitment and Release of Promoter-Proximal RNA Polymerase II. *Molecular Cell* **68,** 491–503.e5. ISSN: 10974164 (2017).

16. Booth, G. T., Parua, P. K., Sansó, M., Fisher, R. P. & Lis, J. T. Cdk9 regulates a promoter-proximal checkpoint to modulate RNA polymerase II elongation rate in fission yeast. *Nature Communications* **9,** 543. ISSN: 2041-1723 (2018).

17. Booth, G. T., Wang, I. X., Cheung, V. G. & Lis, J. T. Divergence of a conserved elongation factor and transcription regulation in budding and fission yeast. *Genome Research* **26,** 799–811. ISSN: 15495469 (2016).

18. Brickner, D. G. *et al.* Subnuclear positioning and interchromosomal clustering of the GAL1-10 locus are controlled by separable, interdependent mechanisms. *Molecular Biology of the Cell* **27,** 2980–2993. ISSN: 1059-1524 (2016).

19. Brown, T. *et al.* Antisense transcription-dependent chromatin signature modulates sense transcript dynamics. *Molecular Systems Biology* **14,** 1–21 (2018).

20. Brown, T., Didelot, X., Wilson, D. J. & Maio, N. D. SimBac: simulation of whole bacterial genomes with homologous recombination. *Microbial Genomics* **2,** 1–6. ISSN: 2057-5858 (2016).

21. Bryant, G. O. *et al.* Activator Control of Nucleosome Occupancy in Activation and Repression of Transcription. **6** (2008).

22. Burger, K. *et al.* 4-Thiouridine inhibits rRNA synthesis and causes a nucleolar stress response. *RNA Biol.* **10,** 1623–1630. ISSN: 15558584 (2013).

23. Burnham, K. & Anderson, D. *Model Selection and Multimodel Inference* 2nd, 70–72. ISBN: 978-0-387-95364-9 (Springer New York, New York, NY, 2004).

24. Candelli, T. *et al.* High-resolution transcription maps reveal the widespread impact of roadblock termination in yeast. *EMBO J.* **37,** e97490. ISSN: 0261-4189 (2018).

25. Carrozza, M. J. *et al.* Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* **123,** 581–592. ISSN: 00928674 (2005).

26. Castelnuovo, M. *et al.* Bimodal expression of PHO84 is modulated by early termination of antisense transcription. *Nature Structural and Molecular Biology* **20,** 851–858. ISSN: 15459993 (2013).

27. Chen, F. X. *et al.* PAF1, a Molecular Regulator of Promoter-Proximal Pausing by RNA Polymerase II. *Cell* **162,** 1003–1015. ISSN: 10974172 (2015).

28. Chen, K. *et al.* DANPOS: Dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Research* **23,** 341–351. ISSN: 10889051 (2013).

29. Choubey, S. Nascent RNA kinetics: Transient and steady state behavior of models of transcription. *Physical Review E* **97,** 022402. ISSN: 2470-0045 (2018).

30. Choubey, S., Kondev, J. & Sanchez, A. Deciphering Transcriptional Dynamics In Vivo by Counting Nascent RNA Molecules. *PLoS Computational Biology* **11,** 1–21. ISSN: 15537358 (2015).

31. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469,** 368–373. ISSN: 1476-4687 (2011).

32. Cloutier, S. C., Wang, S., Ma, W. K., Petell, C. J. & Tran, E. J. Long Noncoding RNAs Promote Transcriptional Poising of Inducible Genes. *PLoS Biology* **11,** 32–34. ISSN: 15449173 (2013).

33. Cloutier, S. C. *et al.* Regulated Formation of lncRNA-DNA Hybrids Enables Faster Transcriptional Induction and Environmental Adaptation. *Molecular Cell* **61,** 393–404. ISSN: 10972765 (2016).

34. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* 2nd ed. ISBN: 9780121790608 (Hillsdale, N.J. ; Hove : Erlbaum Associates, 1988).

35. Collin, P., Jeronimo, C., Poitras, C. & Robert, F. RNA Polymerase II CTD Tyrosine 1 Is Required for Efficient Termination by the Nrd1-Nab3-Sen1 Pathway. *Mol. Cell* **73,** 1–15 (2019).

36. Connelly, S & Manley, J. L. A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. *Genes Dev.* **2,** 440–452. ISSN: 08909369 (1988).

37. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **322,** 1845–1848. ISSN: 0036-8075 (2008).

38. Core, L. J. *et al.* Defining the Status of RNA Polymerase at Promoters. *Cell Reports* **2,** 1025–1035. ISSN: 22111247 (2012).

39. Corrigan, A. M., Tunnacliffe, E., Cannon, D. & Chubb, J. R. A continuum model of transcriptional bursting. *eLife* **5.** ISSN: 2050084X (2016).

40. Courey, A. J. & Jia, S. Transcriptional repression : the long and the short of it. *Genes & Development* **15,** 2786–2796. ISSN: 0890-9369 (2001).

41. Cox, P. M. & Goding, C. R. Transcription and cancer. *British journal of cancer* **63,** 651–62. ISSN: 0007-0920 (1991).

42. Crickard, J. B., Fu, J. & Reese, J. C. Biochemical analysis of yeast suppressor of Ty 4/5 (Spt4/5) reveals the importance of nucleic acid interactions in the prevention of RNA polymerase II arrest. *Journal of Biological Chemistry* **291,** 9853–9870. ISSN: 1083351X (2016).

43. Darzacq, X. *et al.* In vivo dynamics of RNA polymerase II transcription. *Nat. Struct. Mol. Biol.* **14,** 796–806. ISSN: 1545-9993 (2007).

44. Dieppois, G., Iglesias, N. & Stutz, F. Cotranscriptional Recruitment to the mRNA Export Receptor Mex67p Contributes to Nuclear Pore Anchoring of Activated Genes. *Molecular and Cellular Biology* **26,** 7858–7870. ISSN: 0270-7306 (2006).

45. Dobi, K. C. & Winston, F. Analysis of Transcriptional Activation at a Distance in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **27,** 5575–5586. ISSN: 0270-7306 (2007).

46. Duchi, D., Mazumder, A., Malinen, A. M., Ebright, R. H. & Kapanidis, A. N. The RNA polymerase clamp interconverts dynamically among three states and is stabilized in a partly closed state by ppGpp. *Nucleic Acids Res.* 1–12. ISSN: 0305-1048 (2018).

47. Ehrensberger, A. H., Kelly, G. P. & Svejstrup, J. Q. Mechanistic interpretation of promoter-proximal peaks and RNAPII density maps. *Cell* **154,** 713–715. ISSN: 00928674 (2013).

48. Entrevan, M., Schuettengruber, B. & Cavalli, G. Regulation of Genome Architecture and Function by Polycomb Proteins. *Trends in Cell Biology* **26,** 511–525. ISSN: 18793088 (2016).

49. Erickson, B., Sheridan, R. M., Cortazar, M. & Bentley, D. L. Dynamic turnover of paused Pol II complexes at human promoters. *Genes & Development.* ISSN: 0890-9369 (2018).

50. Escalante-Chong, R. *et al.* Galactose metabolic genes in yeast respond to a ratio of galactose and glucose. *Proc. Natl. Acad. Sci.* **112,** 1636–1641. ISSN: 0027-8424 (2015).

51. Eser, P. *et al.* Determinants of RNA metabolism in the Schizosaccharomyces pombe genome. *Mol. Syst. Biol.* **12,** 857–857. ISSN: 1744-4292 (2016).

52. Fan, X. *et al.* Nucleosome depletion at yeast terminators is not intrinsic and can occur by a transcriptional mechanism linked to 3'-end formation. *Proc. Natl. Acad. Sci.* **107,** 17945–17950. ISSN: 0027-8424 (2010).

53. Ferrari, P. & Strubin, M. Uncoupling histone turnover from transcription-associated histone H3 modifications. *Nucleic Acids Research* **43,** 3972–3985. ISSN: 13624962 (2015).

54. Fischl, H., Howe, F. S., Furger, A. & Mellor, J. Paf1 Has Distinct Roles in Transcription Elongation and Differential Transcript Fate. *Molecular Cell* **65,** 685–698.e8. ISSN: 10974164 (2017).

55. Fitz, J., Neumann, T. & Pavri, R. Regulation of RNA polymerase II processivity by Spt5 is restricted to a narrow window during elongation. *The EMBO Journal* **6,** e97965. ISSN: 0261-4189 (2018).

56. Floor, S. N. & Doudna, J. A. Tunable protein synthesis by transcript isoforms in human cells. *eLife* **5.** ISSN: 2050084X (2016).

57. Fu, Y., Pastushok, L. & Xiao, W. DNA damage-induced gene expression in Saccharomyces cerevisiae. *FEMS Microbiology Reviews* **32,** 908–926. ISSN: 01686445 (2008).

58. Garcia, J. F. & Parker, R. MS2 coat proteins bound to yeast mRNAs block 5' to 3' degradation and trap mRNA decay products: implications for the localization of mRNAs by MS2-MCP system. *RNA* **21,** 1393–1395. ISSN: 1355-8382 (2015).

59. Garcia, J. F. & Parker, R. Ubiquitous accumulation of 3' mRNA decay fragments in Saccharomyces cerevisiae mRNAs with chromosomally integrated MS2 arrays. *RNA* **22,** 657–659. ISSN: 14699001 (2016).

60. Gaston, K. & Jayaraman, P. S. Transcriptional repression in eukaryotes: Repressors and repression mechanisms. *Cellular and Molecular Life Sciences* **60,** 721–741. ISSN: 1420682X (2003).

61. Gates, L. A. *et al.* Acetylation on histone H3 lysine 9 mediates a switch from transcription initiation to elongation. *Journal of Biological Chemistry* **292,** 14456–14472. ISSN: 1083351X (2017).

62. Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry* **93555,** 2340–2361. ISSN: 0022-3654 (1977).

63. Ginsburg, D. S. *et al.* NuA4 links methylation of histone H3 lysines 4 and 36 to acetylation of histones H4 and H3. *Journal of Biological Chemistry* **289,** 32656–32670. ISSN: 1083351X (2014).

64. Gressel, S. *et al.* CDK9-dependent RNA polymerase II pausing controls transcription initiation. *eLife* **6,** 1–24. ISSN: 2050-084X (2017).

65. Haimovich, G. *et al.* Gene expression is circular: Factors for mRNA degradation also foster mRNA synthesis. *Cell* **153.** ISSN: 00928674 (2013).

66. Harlen, K. M. & Churchman, L. S. The code and beyond: Transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nature Reviews Molecular Cell Biology* **18,** 263–273. ISSN: 14710080 (2017).

67. Harlen, K. M. *et al.* Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue. *Cell Reports* **15,** 2147–2158. ISSN: 22111247 (2016).

68. Hartzog, G. a., Wada, T, Handa, H & Winston, F. Evidence that Spt4, Spt5, and Spt6 control transcription elongation by RNA polymerase II in *Saccharomyces cerevisiae*. *Genes Dev* **12,** 357–369. ISSN: 0890-9369 (Print) (1998).

69. Heinrich, S., Sidler, C. L., Azzalin, C. M. & Weis, K. Stem-loop RNA labeling can affect nuclear and cytoplasmic mRNA processing. *Rna* **23,** 134–141. ISSN: 14699001 (2017).

70. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A Phase Separation Model for Transcriptional Control. *Cell* **169,** 13–23. ISSN: 10974172 (2017).

71. Hobson, D. J., Wei, W., Steinmetz, L. M. & Svejstrup, J. Q. RNA Polymerase II Collision Interrupts Convergent Transcription. *Molecular Cell* **48,** 365–374. ISSN: 10972765 (2012).

72. Holstege, F. C. *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95,** 717–728. ISSN: 00928674 (1998).

73. Houseley, J. & Tollervey, D. The Many Pathways of RNA Degradation. *Cell* **136,** 763–776. ISSN: 00928674 (2009).

74. Houseley, J., Rubbi, L., Grunstein, M., Tollervey, D. & Vogelauer, M. A ncRNA Modulates Histone Modification and mRNA Induction in the Yeast GAL Gene Cluster. *Mol. Cell* **32,** 685–695. ISSN: 10972765 (2008).

75. Howe, F. S., Fischl, H., Murray, S. C. & Mellor, J. Is H3K4me3 instructive for transcription activation? *BioEssays* **39,** 1–12. ISSN: 15211878 (2017).

76. Huber, F. *et al.* Protein Abundance Control by Non-coding Antisense Transcription. *Cell Reports* **15,** 2625–2636. ISSN: 22111247 (2016).

77. Hughes, A. L., Jin, Y., Rando, O. J. & Struhl, K. A Functional Evolutionary Approach to Identify Determinants of Nucleosome Positioning: A Unifying Model for Establishing the Genome-wide Pattern. *Mol. Cell* **48,** 5–15. ISSN: 10972765 (2012).

78. Hurvich, C. M. & Tsai, C.-L. Regression and Time Series Model Selection in Small Samples. *Biometrika* **76,** 297. ISSN: 00063444 (1989).

79. Ietswaart, R., Rosa, S., Wu, Z., Dean, C. & Howard, M. Cell-Size-Dependent Transcription of FLC and Its Antisense Long Non-coding RNA COOLAIR Explain Cell-to-Cell Expression Variation. *Cell Systems* **4,** 622–635.e9. ISSN: 2405-4712 (2017).

80. Izban, M. G. & Luse, D. S. Factor-stimulated RNA polymerase II transcribes at physiological elongation rates on naked DNA but very poorly on chromatin templates. *Journal of Biological Chemistry* **267,** 13647–13655. ISSN: 00219258 (1992).

81. Jahnke, T. & Huisinga, W. Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of Mathematical Biology* **54,** 1–26. ISSN: 03036812 (2007).

82. Jeronimo, C. & Robert, F. Histone chaperones FACT and Spt6 prevent histone variants from turning into histone deviants. *BioEssays* **38,** 420–426. ISSN: 15211878 (2016).

83. Jin, Y., Eser, U., Struhl, K. & Churchman, L. S. The Ground State and Evolution of Promoter Region Directionality. *Cell* **170,** 889–898.e10. ISSN: 10974172 (2017).

84. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* **316,** 1497–1502. ISSN: 0036-8075 (2007).

85. Karányi, Z. *et al.* Nuclear dynamics of the Set1C subunit Spp1 prepares meiotic recombination sites for break formation. *J Cell Biol,* jcb.201712122. ISSN: 0021-9525 (2018).

86. Katahira, J. MRNA export and the TREX complex. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* **1819,** 507–513. ISSN: 18749399 (2012).

87. Kelleher, R. J., Flanagan, P. M. & Kornberg, R. D. A novel mediator between activator proteins and the RNA polymerase II transcription apparatus. *Cell* **61,** 1209–1215. ISSN: 00928674 (1990).

88. Kent, N. A. & Mellor, J. Chromatin structure snap-shots: Rapid nuclease digestion of chromatin in yeast. *Nucleic Acids Res.* **23,** 3786–3787. ISSN: 03051048 (1995).

89. Keogh, M. C. *et al.* Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. *Cell* **123,** 593–605. ISSN: 00928674 (2005).

90. Kim, J. B. & Sharp, P. A. Positive Transcription Elongation Factor b Phosphorylates hSPT5 and RNA Polymerase II Carboxyl-terminal Domain Independently of Cyclin-dependent Kinase-activating Kinase. *J. Biol. Chem.* **276,** 12317–12323. ISSN: 00219258 (2001).

91. Kim, J. H. *et al.* Modulation of mRNA and lncRNA expression dynamics by the Set2-Rpd3S pathway. *Nature Communications* **7,** 1–11. ISSN: 20411723 (2016).

92. Kim, M. *et al.* The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. *Nature* **432,** 517–522. ISSN: 00280836 (2004).

93. Kim, T., Xu, Z., Clauder-Münster, S., Steinmetz, L. M. & Buratowski, S. Set3 HDAC mediates effects of overlapping noncoding transcription on gene induction kinetics. *Cell* **150,** 1158–1169. ISSN: 00928674 (2012).

94. Knight, B. *et al.* Two distinct promoter architectures centered on dynamic nucleosomes control ribosomal protein gene transcription. *Genes Dev.* **28,** 1695–1709. ISSN: 15495477 (2014).

95. Kuang, Z. *et al.* High-temporal-resolution view of transcription and chromatin states across distinct metabolic states in budding yeast. *Nat. Struct. Mol. Biol.* **21,** 854–863. ISSN: 15459985 (2014).

96. Kubik, S. *et al.* Nucleosome Stability Distinguishes Two Different Promoter Types at All Protein-Coding Genes in Yeast. *Mol. Cell* **60,** 422–434. ISSN: 10974164 (2015).

97. Kumar, N., Singh, A. & Kulkarni, R. V. Transcriptional Bursting in Gene Expression: Analytical Results for General Stochastic Models. *PLoS Computational Biology* **11,** 1–22. ISSN: 15537358 (2015).

98. Kwak, H. & Lis, J. T. Control of Transcriptional Elongation. *Annual Review of Genetics* **47,** 483–508. ISSN: 0066-4197 (2013).

99. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Initiation and Pausing. *Science* **339,** 950–953. ISSN: 0717-6163 (2013).

100. Lafon, A. *et al.* INO80 Chromatin Remodeler Facilitates Release of RNA Polymerase II from Chromatin for Ubiquitin-Mediated Proteasomal Degradation. *Molecular Cell* **60,** 784–796. ISSN: 10974164 (2015).

101. Laitem, C. *et al.* CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II–transcribed genes. *Nature Structural & Molecular Biology* **22,** 396–403. ISSN: 1545-9993 (2015).

102. Larson, M. H. *et al.* A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science (New York, N.Y.)* **344,** 1042–7. ISSN: 1095-9203 (2014).

103. Latchman, D. S. Transcription factors: Bound to activate or repress. *Trends in Biochemical Sciences* **26,** 211–213. ISSN: 09680004 (2001).

104. Lee, B. B. *et al.* Rpd3L HDAC links H3K4me3 to transcriptional repression memory. *Nucleic Acids Research,* 1–14. ISSN: 0305-1048 (2018).

105. Lee, T. I. & Young, R. a. Transcription of Eukaryotic Protein-Coding Genes. *Annu. Rev. Genet.* **34,** 77–137. ISSN: 0066-4197 (2000).

106. Lenstra, T. L., Coulon, A., Chow, C. C. & Larson, D. R. Single-Molecule Imaging Reveals a Switch between Spurious and Functional ncRNA Transcription. *Molecular Cell* **60,** 597–610. ISSN: 10974164 (2015).

107. Li, J. & Zhang, Y. Theoretical analysis of transcription process with polymerase stalling. *Physical Review E* **91,** 052713. ISSN: 1539-3755 (2015).

108. Macqueen, J. *Some methods for classification and analysis of multivariate observations* in *5th Berkeley Symposium on Mathematical Statistics and Probability* (1967), 281–297.

109. Maiuri, P. *et al.* Fast transcription rates of RNA polymerase II in human cells. *EMBO Reports* **12,** 1280–1285. ISSN: 1469221X (2011).

110. Makino, D. L. *et al.* RNA degradation paths in a 12-subunit nuclear exosome complex. *Nature* **524,** 54–58. ISSN: 14764687 (2015).

111. Malik, I., Qiu, C., Snavely, T. & Kaplan, C. D. Wide-ranging and unexpected consequences of altered Pol II catalytic activity in vivo. *Nucleic Acids Research* **45,** 4431–4451. ISSN: 13624962 (2017).

112. Marchler, G, Schüller, C, Adam, G & Ruis, H. A Saccharomyces cerevisiae UAS element controlled by protein kinase A activates transcription in response to a variety of stress conditions. *EMBO J.* **12,** 1997–2003. ISSN: 0261-4189 (1993).

113. Marquardt, S. *et al.* A chromatin-based mechanism for limiting divergent noncoding transcription. *Cell* **157,** 1712–1723. ISSN: 10974172 (2014).

114. Marshall, N. F. & Price, D. H. Purification of P-TEFb, a transcription factor required for the transition into productive elongation. *J. Biol. Chem.* **270,** 12335–12338. ISSN: 00219258 (1995).

115. Mason, P. B. & Struhl, K. Distinction and relationship between elongation rate and processivity of RNA polymerase II in vivo. *Mol. Cell* **17,** 831–840. ISSN: 10972765 (2005).

116. Massey, F. J. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association* **46,** 68–78. ISSN: 0162-1459 (1951).

117. Mayer, A. *et al.* Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* **161,** 541–544. ISSN: 10974172 (2015).

118. Mayer, A., Landry, H. M. & Churchman, L. S. Pause & go: from the discovery of RNA polymerase pausing to its functional implications. *Current Opinion in Cell Biology* **46,** 72–80. ISSN: 18790410 (2017).

119. Mayer, A. *et al.* Uniform transitions of the general RNA polymerase II transcription complex. *Nature Structural & Molecular Biology* **17,** 1272–1278. ISSN: 1545-9993 (2010).

120. Mazza, D., Cella, F., Vicidomini, G., Krol, S. & Diaspro, A. Role of three-dimensional bleach distribution in confocal and two-photon fluorescence recovery after photobleaching experiments. *Applied Optics* **46,** 7401. ISSN: 0003-6935 (2007).

121. Mckay, M. D., Beckman, R. J. & Conover, W. J. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code. *Technometrics* **21,** 239–245 (1979).

122. Medina, D. A. *et al.* Cytoplasmic 5'-3' exonuclease Xrn1p is also a genome-wide transcription factor in yeast. *Frontiers in Genetics* **5,** 1–10. ISSN: 16648021 (2014).

123. Meinhart, A., Kamenski, T., Hoeppner, S., Baumli, S. & Cramer, P. A structural perspective of CTD function. *Genes and Development* **19,** 1401–1415. ISSN: 08909369 (2005).

124. Mellor, J. It Takes a PHD to Read the Histone Code. *Cell* **126,** 22–24. ISSN: 00928674 (2006).

125. Mellor, J., Woloszczuk, R. & Howe, F. S. The Interleaved Genome. *Trends in Genetics* **32,** 57–71. ISSN: 13624555 (2016).

126. Mischo, H. E. & Proudfoot, N. J. Disengaging polymerase: Terminating RNA polymerase II transcription in budding yeast. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1829,** 174–185. ISSN: 18749399 (2013).

127. Mischo, H. E. *et al.* Cell-Cycle Modulation of Transcription Termination Factor Sen1. *Molecular Cell* **70,** 312–326.e7. ISSN: 10974164 (2018).

128. Moretto, F., Wood, N. E., Kelly, G., Doncic, A. & Van Werven, F. J. A regulatory circuit of two lncRNAs and a master regulator directs cell fate in yeast. *Nat. Commun.* **9.** ISSN: 20411723 (2018).

129. Mueller, F., Mazza, D., Stasevich, T. J. & McNally, J. G. FRAP and kinetic modeling in the analysis of nuclear protein dynamics: What do we really know? *Current Opinion in Cell Biology* **22,** 403–411. ISSN: 09550674 (2010).

130. Munchel, S. E., Shultzaberger, R. K., Takizawa, N. & Weis, K. Dynamic profiling of mRNA turnover reveals gene-specific and system-wide regulation of mRNA decay. *Molecular Biology of the Cell* **22,** 2787–2795. ISSN: 1059-1524 (2011).

131. Munsky, B., Li, G., Fox, Z. R., Shepherd, D. P. & Neuert, G. Distribution shapes govern the discovery of predictive models for gene regulation. *Proc. Natl. Acad. Sci.* 201804060. ISSN: 0027-8424 (2018).

132. Murray, S. C. *et al.* A pre-initiation complex at the 3'-end of genes drives antisense transcription independent of divergent sense transcription. *Nucleic Acids Res.* **40,** 2432–2444. ISSN: 1362-4962 (2012).

133. Murray, S. C. *et al.* Sense and antisense transcription are associated with distinct chromatin architectures across genes. *Nucleic Acids Research* **43,** 7823–7837. ISSN: 13624962 (2015).

134. Narita, T. *et al.* Human Transcription Elongation Factor NELF : Identification of Novel Subunits and Reconstitution of the Functionally Active Complex. *Molecular and Cellular Biology* **23,** 1863–1873. ISSN: 0270-7306 (2003).

135. Nehlin, J. O., Carlberg, M & Ronne, H. Control of yeast GAL genes by MIG1 repressor: a transcriptional cascade in the glucose response. *The EMBO journal* **10,** 3373–3377. ISSN: 0261-4189 (1991).

136. Neil, H. *et al.* Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* **457,** 1038–1042. ISSN: 00280836 (2009).

137. Nguyen, T. *et al.* Transcription mediated insulation and interference direct gene cluster expression switches. *eLife* **3,** e03635. ISSN: 2050-084X (2014).

138. Nocetti, N. & Whitehouse, I. Nucleosome repositioning underlies dynamic gene expression. *Genes & Development* **30,** 660–672. ISSN: 0890-9369 (2016).

139. Nojima, T. *et al.* Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* **161,** 526–540. ISSN: 10974172 (2015).

140. Ozcan, S, Vallier, L. G., Flick, J. S., Carlson, M & Johnston, M. Expression of the SUC2 gene of Saccharomyces cerevisiae is induced by low levels of glucose. *Yeast* **13,** 127–137. ISSN: 0749-503X (1997).

141. Palmer, A. C., Egan, J. B. & Shearwin, K. E. Transcriptional interference by RNA polymerase pausing and dislodgement of transcription factors. *Transcription* **2,** 9–14. ISSN: 2154-1264 (2011).

142. Parua, P. K. *et al.* A Cdk9-PP1 switch regulates the elongation-termination transition of RNA polymerase II. *Nature* **558,** 460–464. ISSN: 14764687 (2018).

143. Peccoud, J. & Ycart, B. Markovian modelling of gene product synthesis. *Theoretical Population Biology* **48,** 222–234 (1995).

144. Pei, Y., Schwer, B. & Shuman, S. Interactions between fission yeast Cdk9, its cyclin partner Pch1, and mRNA capping enzyme Pct1 suggest an elongation checkpoint for mRNA quality control. *Journal of Biological Chemistry* **278,** 7180–7188. ISSN: 00219258 (2003).

145. Pelechano, V., Wei, W., Jakob, P. & Steinmetz, L. M. Genome-wide identification of transcript start and end sites by transcript isoform sequencing. *Nature protocols* **9,** 1740–59. ISSN: 1750-2799 (2014).

146. Preker, P. *et al.* RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters. *Science* **322,** 1851–1854. ISSN: 0036-8075 (2008).

147. Prelich, G. RNA polymerase II carboxy-terminal domain kinases: Emerging clues to their function. *Eukaryot. Cell* **1,** 153–162. ISSN: 15359778 (2002).

148. Qiu, Y. & Gilmour, D. S. Identification of regions in the Spt5 subunit of DRB sensitivity-inducing factor (DSIF) that are involved in promoter-proximal pausing. *Journal of Biological Chemistry* **292,** 5555–5570. ISSN: 1083351X (2017).

149. Rabani, M. *et al.* Production and Degradation Dynamics in Mammalian Cells. *Nat Biotechnol.* **29,** 436–442 (2011).

150. Rahl, P. B. *et al.* C-Myc regulates transcriptional pause release. *Cell* **141,** 432–445. ISSN: 00928674 (2010).

151. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biology* **4,** 1707–1719. ISSN: 15457885 (2006).

152. Ravarani, C. N., Chalancon, G., Breker, M., De Groot, N. S. & Babu, M. M. Affinity and competition for TBP are molecular determinants of gene expression noise. *Nature Communications* **7,** 1–13. ISSN: 20411723 (2016).

153. Reinke, H. & Gatfield, D. Genome-wide oscillation of transcription in yeast. *Trends in Biochemical Sciences* **31,** 189–191. ISSN: 09680004 (2006).

154. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147,** 1408–1419. ISSN: 00928674 (2011).

155. Rhee, H. S. & Pugh, B. F. Genome-wide structure and organization of eukaryotic pre- initiation complexes. *Nature* **483,** 295–301 (2012).

156. Riego, L., Avendaño, A., DeLuna, A., Rodríguez, E. & González, A. GDH1 expression is regulated by GLN3, GCN4, and HAP4 under respiratory growth. *Biochemical and Biophysical Research Communications* **293,** 79–85. ISSN: 0006291X (2002).

157. Rodriguez, C. R. *et al.* Kin28, the TFIIH-Associated Carboxy-Terminal Domain Kinase, Facilitates the Recruitment of mRNA Processing Machinery to RNA Polymerase II. *Molecular and Cellular Biology* **20,** 104–112. ISSN: 0270-7306 (2000).

158. Rodríguez-Molina, J. B., Tseng, S. C., Simonett, S. P., Taunton, J. & Ansari, A. Z. Engineered Covalent Inactivation of TFIIH-Kinase Reveals an Elongation Checkpoint and Results in Widespread mRNA Stabilization. *Molecular Cell* **63,** 433–444. ISSN: 10974164 (2016).

159. Rougvie, A. E. & Lis, J. T. The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. *Cell* **54,** 795–804. ISSN: 00928674 (1988).

160. Sansó, M. & Fisher, R. P. Pause, play, repeat: CDKs push RNAP II's buttons. *Transcription* **4,** 146–152. ISSN: 2154-1272 (2013).

161. Santos-Rosa, H. *et al.* Active genes are tri-methylated at K4 of histone H3. *Nature* **419,** 407–411. ISSN: 0028-0836 (2002).

162. Santos-rosa, H. *et al.* Methylation of Histone H3 K4 Mediates Association of the Isw1p ATPase with Chromatin. *Molecular Cell* **12,** 1325–1332 (2003).

163. Saponaro, M. *et al.* RECQL5 controls transcript elongation and suppresses genome instability associated with transcription stress. *Cell* **157,** 1037–1049. ISSN: 10974172 (2014).

164. Schilbach, S. *et al.* Structures of transcription pre-initiation complex with TFIIH and Mediator. *Nature* **551,** 204–209. ISSN: 14764687 (2017).

165. Schlackow, M. *et al.* Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs. *Molecular Cell* **65,** 25–38. ISSN: 10974164 (2017).

166. Schulz, D. *et al.* Transcriptome Surveillance by Selective Termination of Noncoding RNA Synthesis. *Cell* **155,** 1075–1087. ISSN: 00928674 (2013).

167. Schwalb, B. *et al.* TT-seq maps the human transient transcriptome. *Science* **352,** 1225–1228. ISSN: 0036-8075 (2016).

168. Scully, K. H. *et al.* Allosteric effects of Pit-1 DNA sites on long-term repression in cell type specification. *Science* **290,** 1127–1131. ISSN: 00368075 (2000).

169. Segref, A. *et al.* Mex67p, a novel factor for nuclear mRNA export. Binds to both poly(A)+ RNA and nuclear pores. *EMBO Journal* **16,** 3256–3271. ISSN: 02614189 (1997).

170. Sevier, S. A., Kessler, D. A. & Levine, H. Mechanical Bounds to Transcriptional Noise. **113,** 13983–13988. ISSN: 0027-8424 (2016).

171. Shah, N. *et al.* Tyrosine-1 of RNA Polymerase II CTD Controls Global Termination of Gene Transcription in Mammals. *Molecular Cell* **69,** 113–125.e6. ISSN: 10974164 (2018).

172. Shao, W. & Zeitlinger, J. Paused RNA polymerase II inhibits new transcriptional initiation. *Nature Genetics* **49,** 1045–1051. ISSN: 15461718 (2017).

173. Shore, D. & Nasmyth, K. Purification and cloning of a DNA binding protein from yeast that binds to both silencer and activator elements. *Cell* **51,** 721–732 (1987).

174. Sigurdsson, S., Dirac-Svejstrup, A. B. & Svejstrup, J. Q. Evidence that Transcript Cleavage Is Essential for RNA Polymerase II Transcription and Cell Viability. *Molecular Cell* **38,** 202–210. ISSN: 10972765 (2010).

175. Soares, L. M. *et al.* Determinants of Histone H3K4 Methylation Patterns. *Mol. Cell* **68,** 773–785.e6. ISSN: 10974164 (2017).

176. St John, T. P. & Davis, R. W. The organization and transcription of the galactose gene cluster of Saccharomyces. *J. Mol. Biol.* **152,** 285–315. ISSN: 00222836 (1981).

177. Steensels, J. & Verstrepen, K. J. Stop that Noise and Turn Up the Antisense Transcription. *Cell Reports* **15,** 2575–2576. ISSN: 22111247 (2016).

178. Steurer, B. *et al.* Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA Polymerase II. *Proceedings of the National Academy of Sciences* **115,** E4368–E4376. ISSN: 0027-8424 (2018).

179. Strahl, B. D. *et al.* Set2 Is a Nucleosomal Histone H3-Selective Methyltransferase That Mediates Transcriptional Repression. *Molecular and cellular biology* **22,** 1298–1306 (2002).

180. Sun, M. *et al.* Global analysis of Eukaryotic mRNA degradation reveals Xrn1-dependent buffering of transcript levels. *Molecular Cell* **52,** 52–62. ISSN: 10972765 (2013).

181. Svejstrup, J. Q. Contending with transcriptional arrest during RNAPII transcript elongation. *Trends Biochem. Sci.* **32,** 165–171. ISSN: 09680004 (2007).

182. Tan-Wong, S. M. *et al.* Gene Loops Enhance Transcriptional Directionality. *Science* **338,** 671–675. ISSN: 0036-8075 (2012).

183. Tastemel, M. *et al.* Transcription pausing regulates mouse embryonic stem cell differentiation. *Stem Cell Research* **25,** 250–255. ISSN: 18767753 (2017).

184. Tellier, M., Ferrer-Vicens, I. & Murphy, S. The point of no return: The poly(A)-associated elongation checkpoint. *RNA Biology* **13,** 265–271. ISSN: 1547-6286 (2016).

185. Thattai, M. & van Oudenaarden, A. Intrinsic noise in gene regulatory networks. *Proc. Natl. Acad. Sci.* **98,** 8614–8619. ISSN: 0027-8424 (2001).

186. Torchia, T. E., Hamilton, R. W., Cano, C. L. & Hopper, J. E. Disruption of regulatory gene GAL80 in Saccharomyces cerevisiae: effects on carbon-controlled regulation of the galactose/melibiose pathway genes. *Mol. Cell. Biol.* **4,** 1521–1527. ISSN: 0270-7306 (1984).

187. Trcek, T., Larson, D. R., Moldón, A., Query, C. C. & Singer, R. H. Single-Molecule mRNA Decay Measurements Reveal Promoter- Regulated mRNA Stability in Yeast. *Cell* **147,** 1484–1497. ISSN: 00928674 (2011).

188. Trcek, T. *et al.* Single-mRNA counting using fluorescent in situ hybridization in budding yeast. TL - 7. *Nature protocols* **7,** 408–419. ISSN: 1750-2799 (2012).

189. Tschopp, J. F., Emr, S. D., Field, C & Schekman, R. GAL2 codes for a membrane-bound subunit of the galactose permease in Saccharomyces cerevisiae. *J. Bacteriol.* **166,** 313–318. ISSN: 00219193 (1986).

190. Tu, B. P., Kudlicki, A., Rowicka, M. & Mcknight, S. L. Logic of the Yeast Metabolic Cycle : of Cellular Processes. *Science (New York, N.Y.)* **1152,** 1152–8. ISSN: 0036-8075 (2005).

191. Van Dijk, E. L. *et al.* XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature* **475,** 114–117. ISSN: 0028-0836 (2011).

192. Veloso, A. *et al.* Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res.* **24,** 896–905. ISSN: 1088-9051 (2014).

193. Venkatesh, S. *et al.* Set2 methylation of histone H3 lysine 36 suppresses histone exchange on transcribed genes. *Nature* **489,** 452–455. ISSN: 0028-0836 (2012).

194. Venters, B. J. *et al.* A Comprehensive Genomic Binding Map of Gene and Chromatin Regulatory Proteins in Saccharomyces. *Molecular Cell* **41,** 480–492. ISSN: 10972765 (2011).

195. Venturelli, O. S., El-Samad, H. & Murray, R. M. Synergistic dual positive feedback loops established by molecular sequestration generate robust bimodal response. *Proceedings of the National Academy of Sciences* **109,** E3324–E3333. ISSN: 0027-8424 (2012).

196. Vos, S. M. *et al.* Structure of activated transcription complex Pol II–DSIF–PAF–SPT6. *Nature* **560,** 607–612. ISSN: 0028-0836 (2018).

197. Vos, S. M., Farnung, L., Urlaub, H. & Cramer, P. Structure of paused transcription complex Pol II–DSIF–NELF. *Nature* **560,** 601–606. ISSN: 0028-0836 (2018).

198. 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 and Development* **12,** 343–356. ISSN: 08909369 (1998).

199. Wada, T., Takagi, T., Yamaguchi, Y., Watanabe, D. & Handa, H. Evidence that P-TEFb alleviates the negative effect of DSIF on RNA polymerase II-dependent transcription in vitro. *EMBO J.* **17,** 7395–7403. ISSN: 02614189 (1998).

200. Wallace, E. W. & Beggs, J. D. Extremely fast and incredibly close: Cotranscriptional splicing in budding yeast. **23,** 601–610. ISSN: 14699001 (2017).

201. Weinberger, L. *et al.* Expression Noise and Acetylation Profiles Distinguish HDAC Functions. *Molecular Cell* **47,** 193–202. ISSN: 10972765 (2012).

202. Wery, M. *et al.* Nonsense-Mediated Decay Restricts LncRNA Levels in Yeast Unless Blocked by Double-Stranded RNA Structure. *Mol. Cell* **61,** 379–392. ISSN: 10974164 (2016).

203. West, S., Proudfoot, N. J. & Dye, M. J. Molecular Dissection of Mammalian RNA Polymerase II Transcriptional Termination. *Mol. Cell* **29,** 600–610. ISSN: 10972765 (2008).

204. Wu, Z. *et al.* Quantitative regulation of *FLC* via coordinated transcriptional initiation and elongation. *Proceedings of the National Academy of Sciences* **113,** 218–223. ISSN: 0027-8424 (2016).

205. Wyers, F. *et al.* Cryptic Pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121,** 725–737. ISSN: 00928674 (2005).

206. Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457,** 1033–1037. ISSN: 00280836 (2009).

207. Yamada, T. *et al.* P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation. *Molecular Cell* **21,** 227–237. ISSN: 10972765 (2006).

208. Yamaguchi, Y. Y. *et al.* NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell* **97,** 11. ISSN: 00928674 (1999).

209. Yu, M. *et al.* RNA polymerase II – associated factor 1 regulates the release and phosphorylation of paused RNA polymerase II. *Science* **350,** 1383–6 (2015).

210. Zaborowska, J., Egloff, S. & Murphy, S. The pol II CTD: New twists in the tail. *Nature Structural and Molecular Biology* **23,** 771–777. ISSN: 15459985 (2016).

211. Zacharioudakis, I. & Tzamarias, D. Bimodal expression of yeast GAL genes is controlled by a long non-coding RNA and a bifunctional galactokinase. *Biochem. Biophys. Res. Commun.* **486,** 63–69. ISSN: 10902104 (2017).

212. Zaugg, J. B. & Luscombe, N. M. A genomic model of condition-specific nucleosome behavior explains transcriptional activity in yeast. *Genome Research* **22,** 84–94. ISSN: 10889051 (2012).

213. Zenklusen, D., Larson, D. R. & Singer, R. H. Single-RNA counting reveals alternative modes of gene expression in yeast. TL - 15. *Nature structural & molecular biology* **15,** 1263–1271. ISSN: 1545-9985 (2008).

214. Zheng, Q. *et al.* Precise gene deletion and replacement using the CRISPR/Cas9 system in human cells. *Biotechniques* **57,** 115–124. ISSN: 19409818 (2014).

215. Zhou, Q., Li, T. & Price, D. H. RNA Polymerase II Elongation Control. *Annual Review of Biochemistry* **81,** 119–143. ISSN: 0066-4154 (2012).