Transcriptional noise and Pol2 recycling

By
Matthew Jones

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

University of Warwick, School of Life Sciences

April 2022
# Table of Contents

List of Figures 2  
List of Tables 3  
List of Equations 3  
List of Abbreviations 3  
Acknowledgements 6  
Declaration 7  
Publications arising 8  
Abstract 9  
Introduction 10  
Materials and Methods 29  
**Part 1: 3’-5’ crosstalk contributes towards transcriptional noise** 45  
  Chapter 1: An approach with two model genes 45  
  Chapter 2: Measuring 3’-5’ crosstalk and transcriptional noise on a genome wide scale 54  
**Part 2: A novel method to label and track Pol2 from a specific genomic locus** 77  
  Chapter 1: Developing the method 77  
  Chapter 2: Developing a model of transcription to make predictions on polymerase recycling 101  
  Chapter 3: ChIP-qPCR shows polymerase recycling contributes towards transcriptional noise 118  
  Chapter 4: Measuring abortive transcription through a model gene using labelled Pol2 ChIP-Seq 123  
Conclusions 136  
Works cited 143
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Output of Gillespie stochastic simulation</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Schematic of hypothesis</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Kernel density estimates of fluorescence from FLOW-FISH experiment</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Noise plots and parameter estimation from FLOW-FISH experiment</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Noise plots from Influenza infected cells</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Coverage plots at the HBB locus</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Schematic of the ChiA-PET protocol</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Plot of mean expression against 3’-5’ interaction score</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Noise and 3’-5’ interaction plots</td>
</tr>
<tr>
<td>Figure 10</td>
<td>3’-5’ interaction score against transcriptional parameter plots</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Replicate plots of 3’-5’ interactions</td>
</tr>
<tr>
<td>Figure 12</td>
<td>Mean expression and noise plots of looped ribosomal genes</td>
</tr>
<tr>
<td>Figure 13</td>
<td>Mean expression and noise plots of looped and looped ribosomal genes</td>
</tr>
<tr>
<td>Figure 14</td>
<td>Cumulative fraction of nodes against degree plots</td>
</tr>
<tr>
<td>Figure 15</td>
<td>Cumulative fraction of nodes against degree plots fitted to a variety of heavy tailed distributions</td>
</tr>
<tr>
<td>Figure 16</td>
<td>CV^2 against mean plot</td>
</tr>
<tr>
<td>Figure 17</td>
<td>Coverage at TSS plotted against coverage at TES</td>
</tr>
<tr>
<td>Figure 18</td>
<td>ChIP-qPCR of Pol2 occupancy for HBB WT and pA-</td>
</tr>
<tr>
<td>Figure 19</td>
<td>Microscopic model of transcription</td>
</tr>
<tr>
<td>Figure 20</td>
<td>Schematic of experimental protocol</td>
</tr>
<tr>
<td>Figure 21</td>
<td>Western blot for optimising transfection</td>
</tr>
<tr>
<td>Figure 22</td>
<td>Western blots to demonstrate specificity</td>
</tr>
<tr>
<td>Figure 23</td>
<td>Agarose gel to optimise sonication</td>
</tr>
<tr>
<td>Figure 24</td>
<td>Agarose gel of successful library preparation</td>
</tr>
<tr>
<td>Figure 25</td>
<td>CPM normalised metagene plots</td>
</tr>
<tr>
<td>Figure 26</td>
<td>Mitochondrial normalised metagene plots</td>
</tr>
<tr>
<td>Figure 27</td>
<td>Expression of HBB in different cell lines</td>
</tr>
<tr>
<td>Figure 28</td>
<td>Metagenes of anti-biotin ChIP-Seq and Pol2 ChIP-Seq</td>
</tr>
<tr>
<td>Figure 29</td>
<td>Coverage plots of anti-biotin ChIP-Seq and Pol2 ChIP-Seq</td>
</tr>
<tr>
<td>Figure 30</td>
<td>Peak analysis of anti-biotin ChIP-Seq and Pol2 ChIP-Seq</td>
</tr>
<tr>
<td>Figure 31</td>
<td>Boxplots of promoter coverage for genes nearby and far away from the insert</td>
</tr>
<tr>
<td>Figure 32</td>
<td>Schematic of toy model of transcription</td>
</tr>
<tr>
<td>Figure 33</td>
<td>Simulations of model A</td>
</tr>
<tr>
<td>Figure 34</td>
<td>Simulations of model A, demonstrating that this cannot distinguish between recycling and initiation</td>
</tr>
<tr>
<td>Figure 35</td>
<td>Simulations of model B, demonstrating that this model can distinguish between recycling and initiation</td>
</tr>
</tbody>
</table>
List of Tables

| Table 1 | Reagents for Western blot |
| Table 2 | Reagents for ChIP pulldown |
| Table 3 | Reagents for ELISA |
| Table 4 | Reagents for proteomics experiment |
| Table 5 | Reagents for Gibson assembly |
| Table 6 | Reagents for ChIP-qPCR |
| Table 7 | Nodes in/outside hubs overlapping with functional genic elements |
| Table 8 | Predictor P values for linear model |
| Table 9 | Predictor P values for linear model |
| Table 10 | Blacklisted regions |

List of Equations

| Equation 1 | One and two state models of gene expression |
| Equation 2 | Ordinary differential equations governing model B |
| Equation 3 | Ratio of biotinylated Pol2 to total Pol2 |
| Equation 4 | Calculation of the biotinylated Pol2 to total Pol2 ratio |

List of Abbreviations

<p>| DNA       | Deoxyribonucleic acid |
| RNA       | Ribonucleic acid      |</p>
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>rRNA</td>
<td>Ribosomal RNA</td>
</tr>
<tr>
<td>Pol2</td>
<td>RNA Polymerase 2</td>
</tr>
<tr>
<td>GTF</td>
<td>General Transcription factor</td>
</tr>
<tr>
<td>Ser</td>
<td>Serine</td>
</tr>
<tr>
<td>CTD</td>
<td>C-terminal domain</td>
</tr>
<tr>
<td>WT</td>
<td>Wild Type</td>
</tr>
<tr>
<td>pA-</td>
<td>Point mutation in the poly-adenylation signal</td>
</tr>
<tr>
<td>smFISH</td>
<td>Single molecule fluorescence in situ hybridization</td>
</tr>
<tr>
<td>scRNA-Seq</td>
<td>Single cell RNA sequencing</td>
</tr>
<tr>
<td>UMIs</td>
<td>Unique Molecular Identifiers</td>
</tr>
<tr>
<td>TADs</td>
<td>Topologically associating domain</td>
</tr>
<tr>
<td>Hi-C</td>
<td>A method combining chromosome conformation capture with masssively parallel sequencing to probe three-dimensional genome architecture</td>
</tr>
<tr>
<td>ChIA-PET</td>
<td>Chromatin Interaction Analysis by Paired-End Tag Sequencing</td>
</tr>
<tr>
<td>ChIA-Drop</td>
<td>Multiplex chromatin-interaction analysis via droplet-based and barcode-linked sequencing</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>INDELs</td>
<td>Insertion or deletion of bases</td>
</tr>
<tr>
<td>TATA box</td>
<td>Sequence of DNA found in the core promoter region, with a consensus sequence: TATA(A/T)A(A/T)</td>
</tr>
<tr>
<td>CRISPR</td>
<td>Clustered regularly interspaced short palindromic repeats – Used as a genetic engineering technique to edit genomes</td>
</tr>
<tr>
<td>3C</td>
<td>Chromosome conformation capture</td>
</tr>
<tr>
<td>3'</td>
<td>The end of a gene</td>
</tr>
<tr>
<td>5'</td>
<td>The start of a gene</td>
</tr>
<tr>
<td>ChIP-Seq</td>
<td>Chromatin immunoprecipitation sequencing</td>
</tr>
<tr>
<td>TATAlight</td>
<td>A core promoter mutant which is still able to bind TATA binding protein</td>
</tr>
<tr>
<td>TATAmut</td>
<td>A core promoter mutant which is unable to bind TATA binding protein</td>
</tr>
<tr>
<td>TPM</td>
<td>Transcripts per million</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>DMEM</td>
<td>Dulbecco's Modified Eagle Medium</td>
</tr>
<tr>
<td>FBS</td>
<td>Fetal bovine serum</td>
</tr>
<tr>
<td>DMSO</td>
<td>Dimethylsulfoxide</td>
</tr>
<tr>
<td>LN2</td>
<td>Liquid nitrogen</td>
</tr>
<tr>
<td>Avi-RPB1</td>
<td>A fusion protein consisting of the Avitag (Avi), fused to the large subunit of Pol2 (RPB1)</td>
</tr>
<tr>
<td>TetR-BirA</td>
<td>A fusion protein consisting of the Tetracycline repressor protein (TetR) fused to a biotin ligase (BirA)</td>
</tr>
<tr>
<td>PBS</td>
<td>Phosphate-buffered saline</td>
</tr>
<tr>
<td>ChIP-qPCR</td>
<td>Chromatin Immunoprecipitation quantitative real-time PCR</td>
</tr>
<tr>
<td>ABC</td>
<td>Ammonium bicarbonate</td>
</tr>
<tr>
<td>TCEP</td>
<td>tris(2-carboxyethyl)phosphine</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>CAA</td>
<td>Chloroacetamide</td>
</tr>
<tr>
<td>TFA</td>
<td>Trifluoroacetic acid</td>
</tr>
<tr>
<td>ACN</td>
<td>Acetonitrile</td>
</tr>
<tr>
<td>HG38</td>
<td>Homo sapiens (human) genome assembly GRCh38</td>
</tr>
<tr>
<td>MS/MS</td>
<td>Tandem mass spectrometry</td>
</tr>
<tr>
<td>HEK293</td>
<td>Human embryonic kidney cell line</td>
</tr>
<tr>
<td>MCF7</td>
<td>human breast cancer cell line</td>
</tr>
<tr>
<td>K562</td>
<td>Human lymphoblasts cancer cell line</td>
</tr>
<tr>
<td>T-Rex</td>
<td>A cell line stably expressing the Tetracycline repressor protein</td>
</tr>
<tr>
<td>FLOW-FISH</td>
<td>Flow cytometry combined with fluorescent in-situ hybridization</td>
</tr>
<tr>
<td>FISH-QUANT</td>
<td>A program to automatically count the number of transcripts in FISH images</td>
</tr>
<tr>
<td>CV²</td>
<td>Squared coefficient of variation (standard deviation/mean)</td>
</tr>
<tr>
<td>NET-seq</td>
<td>Native elongating transcript sequencing</td>
</tr>
<tr>
<td>A549</td>
<td>A lung cancer cell line</td>
</tr>
<tr>
<td>k_on</td>
<td>Burst frequency</td>
</tr>
<tr>
<td>a/k_off</td>
<td>Burst size</td>
</tr>
<tr>
<td>μ</td>
<td>Mean</td>
</tr>
<tr>
<td>ν</td>
<td>Transcriptional noise, vertical distance from the predicted curve</td>
</tr>
<tr>
<td>HISAT2</td>
<td>hierarchical indexing for spliced alignment of transcripts, a RNA seq alignment software</td>
</tr>
<tr>
<td>IGV</td>
<td>Integrative genomics viewer</td>
</tr>
<tr>
<td>hg19</td>
<td>Homo sapiens (human) genome assembly GRCh37</td>
</tr>
<tr>
<td>GO</td>
<td>Gene ontology</td>
</tr>
<tr>
<td>CDF</td>
<td>Cumulative distribution function</td>
</tr>
<tr>
<td>ChIP-Exo</td>
<td>ChIP-Seq, with an exonuclease digestion stage to improve resolution</td>
</tr>
<tr>
<td>RPKM</td>
<td>Reads Per Kilobase of transcript, per Million mapped reads</td>
</tr>
<tr>
<td>VIF</td>
<td>Variance Inflation Factor</td>
</tr>
<tr>
<td>BioID</td>
<td>Proximity-dependent biotin identification</td>
</tr>
<tr>
<td>Tet</td>
<td>Tetracycline</td>
</tr>
<tr>
<td>CPM</td>
<td>Counts per million</td>
</tr>
<tr>
<td>TSS</td>
<td>Transcriptional start site</td>
</tr>
<tr>
<td>TES</td>
<td>Transcriptional end site</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated region</td>
</tr>
<tr>
<td>CUT&amp;TAG</td>
<td>Cleavage Under Targets and Tagmentation</td>
</tr>
</tbody>
</table>
Acknowledgments

I would like to start by thanking my primary supervisor Daniel Hebenstreit for having such an interesting research topic, which I have really enjoyed investigating for the past four years. I’d like to thank him for giving me freedom to explore ideas, as well as many opportunities to develop myself as a scientist, including peer reviewing articles and collaborations with other research groups.

I’d like to thank all members of the DH research group, past and present for making my time working here so enjoyable, as well as helping me with ideas. There is always a friendly face around ready for a coffee break at a moment’s notice! In particular, I’d like to thank: Massimo Cavallaro, Mark Walsh, Maria Perdiou, Steven Servin and David Edwards.

I’d also like to thank Andrew Bowman, for originally coming up with the idea for the system I developed in Part 2, as well as members of Genomics, Proteomics RTP and Warwick Integrative Synthetic Biology Centre.

This thesis is dedicated to Colin Wadsworth.
Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The work presented (including data generated and data analysis) was carried out by the author except in the cases outlined below:

- smFISH imaging, and data analysis to fit informative priors to FLOW-FISH was mainly carried out by Mark Walsh. I performed smFISH imaging and data analysis on 2 conditions.
- FLOW-FISH experiments were conducted by Mark Walsh and fitted to models of transcription by Massimo Cavallaro.
- UMIs from K562 scRNA-Seq experiment were fitted to models of transcription by Massimo Cavallaro. I adapted these scripts to fit models of transcription to UMIs from scRNA-Seq of influenza infected cells.
- Weina Tang measured the total protein concentration of my cell lines for the ELISA experiment.
- Cleidi Zampronio from the proteomics RTP ran my samples on the mass spectrometer and performed the database searching for the proteomics experiment.

Parts of this thesis have been published by the author:

Massimo Cavallaro, Mark D. Walsh, Matt Jones, James Teahan, Simone Tiberi, Bärbel Finkenstädt & Daniel Hebenstreit
3’-5’ crosstalk contributes to transcriptional bursting
Publications arising

**Peer reviewed:**

Phillip Davies*, **Matt Jones***, Juntai Liu & Daniel Hebenstreit  
*Anti-bias training for (sc)RNA-seq: experimental and computational approaches to improve precision*  
* Authors contributed equally to this work

Monika Sledziowska, **Matt Jones**, Ruba Almaghrabi, Daniel Hebenstreit, Paloma Garcia & Pawel Grzechnik  
*Non-coding RNA Associated With Prader-Willi Syndrome Regulates Transcription Of Neurodevelopmental Genes In Human Induced Pluripotent Stem Cells*  

**In review:**

Kinga Winczura, Hurmuz Ceylan, Monika Sledziowska, **Matt Jones**, Holly Fagarasan, Jianming Wang, Marco Saponaro, Roland Arnold, Daniel Hebenstreit, Pawel Grzechnik  
*RPRD Proteins Control Transcription in Human Cells*  
Abstract

Isogenic cells in identical environments can have a high variation of mRNA transcribed from a particular gene within their population, often exceeding expectations based on a constant transcription rate. This cell-to-cell variation is called transcriptional noise. This thesis aims to investigate the hypothesis that one origin of transcriptional noise is crosstalk between 3’ and 5’ ends of genes, with the proposed mechanism that this crosstalk facilitates the local recycling of RNA polymerase 2. I begin by describing a study of transcriptional noise on two model genes and combine this with a genome wide approach to demonstrate that crosstalk contributes to transcriptional noise. I then develop a novel technology to label RNA polymerase 2 at a specific genetic location, which I then use to measure polymerase recycling to experimentally verify the proposed mechanism. I demonstrate on a model gene that polymerase recycling is reduced upon ablation of 3’-5’ crosstalk. I then use this technology to measure abortive transcription throughout a model gene. I demonstrate that abortive transcription rates reduce as RNA polymerase 2 progresses through the model gene.
Introduction

Transcription

Transcription is the process of copying the genetic information carried in a segment of DNA into RNA. This process is fundamental for all life, as DNA by itself is a carrier of genetic information but is non-functional in natural systems. RNA can be functional by itself, due to its propensity to form secondary structures, with an important example being components of the ribosome. Some RNA, termed messenger RNA (mRNA) is translated by the ribosome to make a different type of functional product – a protein, which performs the majority of cellular functions.

The enzymes which perform transcription are called RNA polymerases. These are multi-subunit enzymes which move stepwise along DNA and catalyse the extension of an RNA nucleotide chain. In eukaryotes, RNA polymerase 2 (Pol2) is the enzyme which catalyses the production of mRNA, which is translated to create proteins. This thesis will be concentrating on transcription by the enzyme Pol2.

Transcription is a process that can be broadly separated into three stages: initiation, productive elongation, and termination. I will briefly describe each of these stages in turn, focusing on transcription of human protein coding genes by Pol2. In initiation, a common sequential model states a general transcription factor TFIID binds to a core promoter motif and recruits many general transcription factors (GTF) such as TFIIA and TFIIB. These GTFs recruit Pol2 to form the pre-initiation complex (Haberle and Stark, 2018). Distal enhancer elements can promote this assembly by recruiting transcription factors and cofactors. One member of the pre-initiation complex, TFIIH, has two major roles. TFIIH phosphorylates serine-5 (ser-5) of the heptad repeat (consensus sequence YSPTSPS, ser-5 underlined) on the C-terminal domain (CTD) of the largest subunit of Pol2. TFIIH also acts to unwind DNA (Compe and Egly, 2012). This allows transcription to be initiated.
Pol2 is thought to pause approximately 30-50 nucleotides after transcription is initiated, as high polymerase density is often found at transcriptional start sites. However, this has been challenged recently as abortive transcription may also explain the high density of polymerases at transcriptional start sites (Erickson et al., 2018; Krebs et al., 2017; Steurer et al., 2018; Zhang et al., 2021). Nevertheless, Pol2 escapes polymerase pausing/abortive transcription through phosphorylation of the ser-2 on the CTD along with factors such as the proteins NELF and DSIF, causing dissociation of these factors and allowing Pol2 to enter productive elongation (Haberle and Stark, 2018).

After productive elongation, transcription is terminated. Here the CPSF-CF complex binds to the ser-2 phosphorylated CTD of Pol2, and sequences such as the poly-adenylation signal are recognized causing cleavage of the mRNA (Porrua and Libri, 2015). Two models have been proposed to explain how Pol2 subsequently unbinds the RNA strand; the allosteric model suggests that Pol2 detects progression through the poly-adenylation signal and releases the RNA strand. The torpedo model, on the other hand, posits that the 5’ to 3’ exonuclease Xrn2 degrades the downstream transcript until the still elongating Pol2 is reached. Arrival of Xrn2 to Pol2 then helps to release Pol2 from the DNA strand (Proudfoot, 2016). Indeed, depletion of Xrn2 has been shown to reduce termination (West et al., 2004), providing evidence for the torpedo model.

Transcription is a key stage of gene regulation, where a cell can increase or decrease the production of mRNA for a gene based on some stimuli, which will then result in changes in the concentrations of proteins encoded by mRNA. Classical examples of these include signalling cascades, where a ligand binding event on the cell surface starts a phosphorylation cascade, resulting in a translocation of an enzyme to the nucleus and activating a host of transcription factors. These bind specific DNA sequences to repress or activate expression of given genes (e.g., the mitogen-activated protein kinase signalling cascade). Multiple other mechanisms of regulation exist such as chromatin remodelling, histone modifications and long-range interactions.
Even with this high degree of regulation, isogenic cells in identical environments can have considerable variation in the numbers of mRNA transcribed by a given gene. This is termed transcriptional noise.

Transcriptional noise

There are two sources of noise in transcription, extrinsic and intrinsic ones. Extrinsic sources derive from cell-to-cell differences in external factors, for example cell size, cell cycle stage, and differences in local environments (e.g., number of surrounding cells, nutrient availability, etc) which can cause cell-to-cell variation in gene expression. Intrinsic sources however are due to the inherent stochasticity of biological systems. Both extrinsic and intrinsic sources of noise can be demonstrated by engineering expression of two fluorescent reporter proteins within the same cell, which are controlled by the same promoter (Elowitz et al., 2002). In this system, within a cell there is no difference in external factors between these two genes, so differences between fluorescence intensity between both reporters is likely to be due to intrinsic noise, and fluctuations of these reporters in a correlated fashion are due to extrinsic noise.

As previously mentioned, intrinsic noise is a consequence of the inherent stochasticity in biological systems. For example, every reaction that occurs requires randomly diffusing reactants to encounter each other, often in the presence of another randomly diffusing enzyme. As there are low numbers of biological components present in cells (e.g. two copies of most genes) this increases the noise (Gillespie, 1976).

Measuring the distribution of numbers of mRNA transcribed by a particular gene, from a population of cells in identical environments can give information on how this gene is transcribed. For example, the distribution of mRNA for some genes follows a Poisson distribution, where the variance equals the mean expression. A model which can replicate this distribution is a gene which is transcribed at a constant - though
stochastic - rate (one-state model Equation 1 A, See Figure 1 A for a simulation of this model) (Munsky et al., 2012).

This one state model seems to describe transcription of many genes well; droplet based scRNA-Seq suggests that the distributions of mRNA transcribed by many individual genes are consistent with Poisson statistics (Klein et al., 2015) and imaging studies show housekeeping genes in budding yeast are also well described by the Poisson distribution (Zenklusen et al., 2008). However, for some genes the expression distribution is wider than the Poisson, with the variance being above mean expression. A simple model which can replicate this distribution can be described by a gene having two states (Equation 1 A), on and off, with the gene only producing mRNA in the on state (Peccoud and Ycart, 1995). See Figure 1 B for a simulation of this model.

A  
One – state model: Gene → mRNA → degraded mRNA

Two – state model: Gene off ↔ Gene on → mRNA → degraded mRNA

B  
RNA | k\text{transcribe},p \sim \text{Poisson}(k\text{transcribe} \times p)
P | \k\text{on},\k\text{off} \sim \text{Beta}(\k\text{on},\k\text{off})

Equation 1. (A) One and two state models of gene expression. (B) Poisson Beta mixture distribution describing the two-state model of gene expression, here p is an auxiliary variable which is described by a beta distribution of the rates of a gene moving to the on and off states.

The two-state model predicts ‘bursts’ of transcription (Figure 1 B), where many mRNAs are made when the gene is in the ‘on’ state. Observations from live-cell imaging experiments match the predictions of transcriptional bursting from the two-state model (e.g. Chubb et al., 2006). Stationary states of mRNA from a cell population can be fitted to the Poisson beta mixture distribution (Equation 1 B, Figure 1 C), which describes this model, (Kim and Marioni, 2013) thus allowing estimations of transcriptional parameters. Key parameters to be inferred is the burst frequency
(rate of a gene moving into the on state), and the burst size, defined as the average number of mRNAs produced per burst.

Due to the principles of parsimony, this two-state model is the most prevalent model used to describe transcriptional noise, especially when measuring steady state mRNA distributions. However other models including a gene having the ability to exist in multiple states have been proposed, which in some cases can match live imaging data better than the two-state model (Tunnacliffe and Chubb, 2020).

Figure 1 An output from a Gillespie stochastic simulation implemented with the R package GillespieSSA (Pineda-Krch, 2008), used to demonstrate transcriptional bursting. Both the one-state and the two-state models have the same mean mRNA expression. (A) Simulation results from the one-state model, tracking the number of mRNAs through time. Here, the mRNA levels stay near the mean value. (B) Simulation results from the two-state model, tracking the number of mRNAs through time. Here, mRNA is produced in bursts. (C) A histogram showing the density of mRNA numbers across these simulations. Here, the two-state model has more transcriptional noise than the one-state model (a wider mRNA distribution) with the same mean mRNA expression.
Function of noise

Transcriptional noise between genetically identical cells can cause cell-to-cell variation in protein concentrations, which can propagate to downstream processes (Eldar and Elowitz, 2010). Some classes of genes have higher variabilities in protein concentrations than others (Newman et al., 2006), which suggests transcriptional noise is evolvable (Lehner, 2008). In some cases this variability seems to have been selected for; gene expression variation of AcrAB-TolC (amulti drug efflux pump) in *E. coli* can result in some cells being able to survive antibiotic treatment, with a trade-off of reduced growth rates (El Meouche and Dunlop, 2018). This could be interpreted as a ‘bet-hedging strategy’ suggesting that evolution of this trait has occurred due to fluctuating environments (Simons, 2011). Indeed, bet-hedging strategies have been shown to evolve through experimental evolution in fluctuating environments (Beaumont et al., 2009). In mammalian systems, an example of functional cell-to-cell variability is the probabilistic differentiation of the inner cell mass of the early mouse embryo into primitive endoderm and epiblast lineages, likely caused by transcriptional noise (Dietrich and Hiiragi, 2007).

Measuring Transcription

To understand transcriptional noise, it is important to accurately measure mRNA abundances inside single cells. There are multiple approaches which can be used to do so. These approaches can naturally be split into two groups. In the first approach, cells are alive and transcription can be measured for an extended period of time with transcription and transcriptional fluctuations being directly observed. The other approach is where a snapshot of many fixed cells is captured, and transcriptional noise is inferred by the distribution of mRNA.

The first approach utilises live cell imaging to measure transcription. Here, the gene coding for an mRNA of interest is edited to contain MS2 or PP7 stem loop motifs (Bertrand et al., 1998; Larson et al., 2011). Viral coat proteins fused to a fluorescent reporter protein which bind to the stem loops are expressed. This allows visualisation of nascent mRNA as a dot on the locus that is being expressed (Chubb et al., 2006).
Parameters such as fluorescence intensity of the dot (proportional to total numbers of mRNAs being produced), time between dots, and time that the dot is expressed for allow parameters such as burst size, burst frequency, and burst duration to be estimated.

Whilst this approach can directly visualise transcription in living cells, it does require genetic engineering to insert these stem loop motifs, which is time consuming and potentially changes the endogenous behaviour of the gene. As there are few well-defined viral coat proteins that bind to different stem loop motifs few RNAs can be measured simultaneously. It is also low throughput, like many imaging-based techniques.

Another imaging-based approach to measure transcriptional noise is to measure mRNA from many fixed cells. The distribution of mRNA is then used to infer transcriptional parameters by fitting the results to a model. Single molecule RNA-FISH (smFISH) (Femino et al., 1998; Raj et al., 2008) where each transcript appears as a single diffraction limited spot allows absolute quantification of the numbers of mRNA molecules by counting the number of dots. Software tools have been developed to assist this (Mueller et al., 2013). Again, due to smFISH being an imaging-based technique, it has low throughput. Imaging based smFISH can be combined with flow cytometry to estimate the number of mRNAs from overall fluorescence, allowing the mRNA numbers of tens of thousands of cells to be estimated (Cavallaro et al., 2021; Tiberi et al., 2018), thus helping improve the throughput.

One problem with RNA-FISH based techniques is that only a few genes are available to detect simultaneously as there is only a small number of non-overlapping fluorophores available. Temporal barcoding techniques have been developed (Chen et al., 2015; Lubeck et al., 2014; Shah et al., 2016) allowing hundreds, and more recently ten thousand (Eng et al., 2019) mRNAs to be imaged and counted. However, this technique is laborious and expensive – making it currently impossible for many laboratories to measure the whole transcriptome with RNA FISH based methods.
Another method used to measure the transcriptome in single cells, through taking a snapshot based approach, is single-cell RNA sequencing (scRNA-Seq) (Tang et al., 2009). Here the transcriptome is converted into cDNA, amplified, and sequenced. This is cheaper and has higher throughput than FISH based techniques, with many commercial implementations. Indeed, the number of cells being sequenced is scaling exponentially over time (Svensson et al., 2018). A key innovation which allowed this is some way of separating individual cells in droplets (Klein et al., 2015) so that barcodes can be added and reads can be assigned to cells bioinformatically. Unique molecular identifiers (UMIs) (Islam et al., 2014; Kivioja et al., 2012) are often combined with the barcoding approach which (if assumed to be completely efficient) will result in every mRNA being counted only once. We have successfully used this approach (Part 1, Chapter 2, (Cavallaro et al., 2021)) to estimate transcriptional noise using deposited data (Klein et al., 2015).

Unfortunately, scRNA-Seq is not fully efficient, with estimates of capture efficiency ranging from 10-40% (Haque et al., 2017), resulting in many expressed transcripts being lost, termed dropouts. This is thought to result in zero-inflated datasets, where many cells unexpectedly have no reads for genes. However, benchmarking datasets, show that the number of zero values is consistent with expectations from count data (Svensson, 2020). Other sources of errors in scRNA-sequencing datasets include library preparation efficiency, transcriptome size differences, batch effects, doublets and ambient gene expression, which are described in Davies et al. (2021).

Often, the results from imaging based studies such as smFISH -which are thought to be the gold standard- and scRNA-Seq, do not exactly match each other (Huang et al., 2018). In particular, genes appear to show less cell-to-cell variation in data obtained from scRNA-Seq compared to smFISH (Davies et al., 2021; Huang et al., 2018). Software tools have been developed (e.g. Huang et al., 2018; Tang et al., 2020) which aim to impute more realistic values of gene expression from scRNA UMI counts, and these distributions seem to match RNA FISH well. However, these approaches may also introduce artifacts to scRNA-Seq data (Davies et al., 2021). Even with these
issues, scRNA-Seq is a very powerful technique for measuring transcriptional noise on a genome wide scale.

In this thesis, I utilise the snapshot approach. I performed smFISH experiments as part of a collaborative effort to measure transcriptional noise at a certain gene (Cavallaro et al., 2021) which I will briefly describe in this thesis. I also used a pre-existing scRNA-Seq dataset (Klein et al., 2015) in collaboration with Massimo Cavallaro to estimate transcriptional noise at a genome wide scale (Cavallaro et al., 2021).

I also develop my own technology for a bulk population of cells. This technology holds, labels and releases Pol2 from a specific location in the genome. I then isolate labelled Pol2 after release and can sequence DNA associated with labelled Pol2. This can be used to track the diffusion of labelled Pol2 which I use to measure both polymerase recycling and abortive transcription. This technique is currently used for a bulk population of cells, averaging the results over millions of cells, but is in principle adaptable to single cell resolution.

The three-dimensional genome

The cell nucleus is approximately 8 µm in diameter in mammalian cells, which appears hard to reconcile with the approximate length of 2m of DNA which needs to be packed into each nucleus, a 250,000-fold condensation. This DNA is not packaged randomly. For example, early microscopy experiments staining chromatin showed that some areas of chromatin are more dense than others which suggests that the cell has some form of control of where DNA is packed (Heitz, 1928).

The basic unit of packing is the nucleosome (Khorasanizadeh, 2004). This is an octameric protein consisting of four core histone proteins H2A, H2B, H3 and H4, though variants of these exist (Kamakaka and Biggins, 2005). DNA is wrapped twice around the 11nm diameter of this turn which consists of approximately 146 base pairs (Olins and Olins, 2003), with a linker region of around 60 bp until the next
nucleosome. This results in a ~10 nm fibre of nucleosomes which appear as ‘beads on a string’ in electron microscopy (Olins and Olins, 2003), however the structure of this fibre has a varying diameter and is likely heterogeneous (Maeshima et al., 2019). DNA can also loop, placing distal DNA regions in close three-dimensional proximity. These loops can be measured on a genome wide scale (e.g. Rao et al., 2014) and their anchors have been found to be enriched in cohesin and CTCF in a convergent orientation. Loop extrusion is thought to be a mechanism through which these loops are formed (Sanborn et al., 2015).

Other interactions, such as longer-range folding, occur too. One example of this is topologically associating domains (TADs). These are self-interacting regions of the genome, where DNA sequences inside a TAD are in close three-dimensional proximity to each other (Pombo and Dillon, 2015). Larger contact domains also exist, termed A and B (Lieberman-Aiden et al., 2009) which will be described in the following section.

There are a few methods used for measuring the three-dimensional genome, generally split into imaging approaches (live or fixed cells) which can measure the three-dimensional genome on a single cell level, and sequencing approaches which provide a higher resolution but measure the average interactions of (generally) millions of cells.

Imaging approaches commonly use DNA-FISH, where fluorescently labelled DNA probes bind to a region of interest or DNA binding proteins fused to a fluorophore are used to visualise live cells (Kempfer and Pombo, 2020). The distance between the coloured dots can then be measured to give a meaningful absolute distance between genomic loci. The resolution of light microscopy is fairly low however, and super resolution techniques have been developed which give resolution on the order of tens of nanometres (Boettiger et al., 2016). Generally, these approaches are limited to a few loci at once due to the unavailability of many non-overlapping fluorophores.
Sequencing based approaches generally depend on proximity ligation. Here, the DNA is fixed, sheared and ligated back together, and fragments which are closer together in three-dimensional space are more likely to do so. Paired-end sequencing is then used to build up contact matrices of interactions. One of these approaches (Hi-C) gives higher resolution than imaging, and on a genome-wide scale. For example Bonev et al. 2017 achieved a resolution of 750 bp. Other approaches such as ChIA-PET (Fullwood et al., 2009) take a similar approach but with a chromatin enrichment stage allowing chromatin structure to be measured in the proximity to proteins of interest.

How does the three-dimensional genome affect transcription?

The first Hi-C approaches discovered chromatin was separated into two major domains, termed A and B (Lieberman-Aiden et al., 2009). Sequences in each component were found to preferentially interact with other sequences in the same component. The A component is enriched in transcribed genes and activating histone marks, whereas genes in the B component are generally silenced (Lieberman-Aiden et al., 2009). These components appear to represent euchromatin and heterochromatin visible through microscopy (Heitz, 1928).

Other aspects of the three-dimensional genome also affect transcription. One example is enhancer elements which increase the transcriptional output of target genes many kilobases, or even megabases away from each other (Krivega and Dean, 2012). It seems likely that the only explanation for this is for chromatin to fold in such a way to bring enhancer elements towards promoters to enhance transcription. One imaging based experiment from Carter et al. (2002) demonstrates this well; here the authors found that the distal locus control region of the HBB gene (over 50 kb away from the gene) is in close three-dimensional proximity to HBB. Furthermore Hi-C based approaches confirm this on a genome wide scale, as they have found that more loops than expected by chance are anchored on either end by known enhancers and promoters (Rao et al., 2014).
Imaging approaches suggest that these enhancer-promoter interactions are not simply one to one. As in, a single enhancer does not necessarily interact with a single promoter. Some genes controlled by the same enhancer show co-ordinated transcription contrary to the classical model of enhancers looping towards 1 promoter at a time (Fukaya et al., 2016), which can even happen when genes are on different chromosomes (Lim et al., 2018). This may suggest a more complex interplay of enhancer-promoter interactions, with potentially multiple distal DNA regions all coming together simultaneously. This hypothesis cannot be tested with Hi-C or ChIA-PET based approaches alone, as these can only measure pairwise interactions.

Combining microfluidic droplet-based barcoding with fragmented fixed chromatin results in spatially close chromatin residing in the same droplet. This technique can therefore allow multiplexed interactions to be discovered, and is known as ChIA-Drop (Zheng et al., 2019). Approximately half of all complexes contain more than two fragments which supports the hypothesis that multiple distal DNA regions can come together simultaneously, and interactions are more complex than pairwise interactions measured in Hi-C and ChIA-PET. The authors of the ChIA-Drop study also found that upon enriching chromatin complexes for Pol2, a large proportion of these chromatin complexes (20%) contain multiple promoters, suggesting these complex interactions of multiple DNA regions are transcriptionally relevant.

Other work, using live cell imaging with MS2, PP7 and ParB (DNA binding protein) showed that chromatin is dynamic and that proximity to a distal enhancer (~150 kb) is required for some genes’ activation (Chen et al., 2018). However, live imaging studies showed that while the Sox2 enhancer is required for transcription of Sox2, its transcriptional activity is not correlated with the gene’s proximity to its enhancer (Alexander et al., 2019), suggesting there may not be a universal mechanism for enhancers.

As with other imaging approaches mentioned, these approaches for measuring the three-dimensional genome structure have low throughput, with only a few loci being able to be simultaneously measured. Furthermore, their resolution is too low to
investigate local three-dimensional interactions. In this thesis, I analyse a publicly available ChIA-PET dataset (Li et al., 2012) to measure crosstalk between the end and start of a gene (3'-5' crosstalk) on a genome wide scale.

**Known contributors to transcriptional bursting**

It is not fully understood how transcriptional bursting arises; however, it has been found across a wide diversity of systems including viruses, prokaryotes and eukaryotes (Hebenstreit, 2013).

It is known that the local DNA sequence around the gene has a role. For example, the amoeba *D. discoideum* encodes 17 Actin genes, all encoding the same protein. The bursting dynamics of all these genes are different, and switching the promoter regions of actin genes results in bursting dynamics also switching (Tunnacliffe et al., 2018).

A recent, genome wide method to estimate intrinsic noise in eukaryotic cells follows a conceptually similar approach to Elowitz et al. (2002) and is based on single cell RNA sequencing (scRNA-Seq) (Ochiai et al., 2020). Here, two different inbred mouse strains are crossed, and scRNA-Seq is performed on the resulting embryonic stem cells, with single nucleotide polymorphisms (SNP) differences and insertion or deletion of bases (INDELs) between the two inbred lines being used to determine which alleles are being expressed. Assuming that the majority of these SNPs and INDELs are non-functional (Kimura, 1968), this means that differences between the expression of two alleles the same cell is due to intrinsic noise. Taking the same method as Elowitz et al. (2002) and knocking in dual fluorescent reporters on 25 targets the authors confirmed correlation of intrinsic noise from scRNA-Seq to measured intrinsic noise on the protein level (Ochiai et al., 2020). Here, it was also found that the promoter region contributes towards the regulation of transcriptional bursts, as the presence of a TATA box as well as other factors including EP300, ELL2 and MED12 in the promoter correlated with burst size (Ochiai et al., 2020). Other
experiments also taking a similar approach using allele sensitive scRNA-seq also find the presence of a TATA box contributes to burst size (Larsson et al., 2019).

An imaging based study (Pimmett et al., 2021) also shows the presence of a TATA box which is able to bind TBP contributes to burst sizes, supporting the results from the sequencing based studies described above. Here, the authors mutate the promoter sequence of a model minigene containing 24x MS2 stem loops and a yellow fluorescent reporter in drosophila embryos. They use this to image transcription with different promoter mutations in the same genomic context, to see how transcriptional bursting changes upon different promoters. The TATAlight promoter (still able to bind TBP) and TATAmut promoter (unable to bind TBP) appear to have similar burst frequencies, but TATAmut has much lower burst sizes, and appears to have less transcriptional noise than TATAlight.

Distal Enhancer activity has also been shown to play a role in transcriptional bursting. Imaging studies have demonstrated that inhibiting enhancer activity through the addition of insulator elements seem to reduce the burst frequency, rather than the burst size (Fukaya et al., 2016), which was also supported by scRNA-seq experiments (Larsson et al., 2019).

A lentiviral CRISPR knockout screen, combined with scRNA-seq, has also been conducted in order to discover genes which contribute towards transcriptional noise (Ochiai et al., 2020). Here, the authors found that genes in the mTOR and MAPK signalling pathways were involved at promoting transcriptional noise, which was confirmed through drug-based inhibition of these pathways.

Work in our laboratory which I have contributed to and will discuss further in this thesis, has also shown that mutating the poly-adenylation signal at the 3’ end of genes can contribute to transcriptional bursting (Cavallaro et al., 2021).
Abortive transcription

Abortive transcription is an understudied process, where in transcription Pol2 unbinds from the DNA strand, and stops active transcription. Early work describing this effect was focused on characterising this effect in prokaryotic initiation. This involved high resolution gel electrophoresis in vitro (Carpousis and Gralla, 1980) and in vivo (Goldman et al., 2009), where short RNA strands (2-15 nucleotides) are synthesised and released. Indeed, a study found between 7, 32 and 165 abortive transcripts were made per full length transcript, depending on the promoter measured (Hsu et al., 2003).

More recent work has suggested that abortive transcription is common amongst eukaryotes too, and mainly occurs towards the start of transcription. As mentioned earlier, the high polymerase density at transcriptional start sites, often interpreted as polymerase pausing, could actually be caused by a high degree of abortive transcription at these regions (Erickson et al., 2018; Krebs et al., 2017; Steurer et al., 2018; Zhang et al., 2021). Live cell imaging combined with drug treatments to block CDK7 (a subunit of TFIIH, which phosphorylates Ser-5 on Pol2 CTD and facilitates initiation) and CDK9 (which phosphorylates Ser-2 on Pol2 CTD, NELF and DISF; facilitating entry to productive elongation) showed high rates of abortive transcription in eukaryotic cells too (Steurer et al., 2018). The authors found 87.3% of transcripts abort in initiation and 92.4% of the remaining transcripts before productive elongation. Other work also finds a high degree of Pol2 turnover at promoters (Erickson et al., 2018; Krebs et al., 2017).

Abortive transcription has also been measured in productive elongation. For example, mutations in Pol2 subunits, elongation factors and drug treatments appear to increase abortive transcription, measured by comparing Pol2 density across a gene to a control without the mutation or treatments (e.g. Mason and Struhl, 2005). A potential issue with this approach, however, is that polymerase density is composed of both polymerase pausing, and polymerase flux (Zhang et al., 2021). Whilst abortive transcription will reduce polymerase flux and reduce polymerase density, changes in
polymerase pausing across a gene will also contribute to changes in polymerase density across a gene. This makes measuring abortive transcription by measuring polymerase density alone difficult.

I use the technique I develop in Part 2, Chapter 4 to measure relative rates of abortive transcription throughout a model gene. This technique uses a separate approach from measuring polymerase density throughout the gene and is less sensitive to pausing.

**Aims and open questions**

My thesis will explore how the three-dimensional genome affects transcription on a more local scale. I will start by testing the hypothesis that close three-dimensional proximity between both sides of a gene (3’-5’ crosstalk) contributes to transcriptional noise (Hebenstreit, 2013). I will then test the proposed mechanism that local recycling of Pol2 causes this effect (Figure 2). The rationale behind this hypothesis and mechanism is that 3’-5’ crosstalk will allow fast diffusion from the end of the gene back to the start, permitting the same Pol2 to re-initiate transcription rather than diffusing to another gene. This could result in Pol2 transcribing ‘in circles’, producing many mRNAs in a burst until this crosstalk is disrupted or other transcription factors dissociate from the gene (Hebenstreit, 2013).
There is plenty of evidence that points towards this hypothesis, including, in particular, experiments looking at mutations in polyadenylation sites at the 3’ end of a gene (Mapendano et al., 2010). Here the authors integrated single copies of beta-globin (HBB) or HIV env genes, under the control of the Tet operon into HEK293 cell lines. These integrated genes either have wild type (WT) poly-adenylation signal, or a point mutation A₂TA₃ -> A₂GA₃ (pA⁻). Crucially, after tetracycline induction, this point mutation (3’ end) resulted in reduced Pol2 occupancy at the promoter (5’ end). This provides evidence for crosstalk between the start and the end of a gene and suggests that the point mutation reduces re-initiation of transcription.

Figure 2. A schematic of the hypothesis I am testing in this thesis. (A) represents a gene with 3’-5’ crosstalk, the transcriptional start site and the transcriptional end site are in close three-dimensional proximity. This means Pol2 can quickly re-initiate transcription, transcribing in circles and causing a burst of transcription, resulting in increased transcriptional noise. (B) represents a gene without 3’-5’ crosstalk. Here, the transcriptional start site and transcriptional end site are far apart, so Pol2 does not re-initiate transcription resulting in this gene having lower transcriptional noise.
Later work (Tan-Wong et al., 2012) used 3C techniques to show that the 3’ end and 5’ end of HBB were in close three-dimensional proximity with the WT polyadenylation signal but this was disrupted upon the point mutation. Again, this provides further evidence for crosstalk between the start and end of a gene, which is disrupted upon polyadenylation site mutations.

Other evidence points towards 3’-5’ crosstalk, for example ChIP-Seq datasets show that the poly-adenylation factor PCF11 is located at both the 5’ and the 3’ end of genes (Kamieniarz-Gdula et al., 2019). Many other factors, for example TFIIB, a common transcription initiation factor, are involved in processes at both the 5’ and 3’ end of a gene. TFIIB can bind to factors yCPF, hCPSF and CstF all of which are involved in transcriptional termination (Shandilya and Roberts, 2012). Indeed, depletion of TFIIB has also directly been shown to reduce 3’-5’ proximity, shown through 3C techniques on a series of model genes, though this was conducted in yeast and not a mammalian system (Singh and Hampsey, 2007).

As part of my thesis I have contributed to some published research which used the HBB and HIV env cell lines described above (Cavallaro et al., 2021) and showed that this crosstalk is transcriptionally relevant and contributes towards transcriptional noise. For the gene HBB, the mutant lines had lower transcriptional noise, higher burst frequencies, and lower burst sizes at the same overall expression level. The HIV env lines were less conclusive but supported the results.

It is important to note however, that 3’-5’ crosstalk cannot be the sole cause of transcriptional noise. A recent study based on fluorescent labelling and imaging of long genes (≥100 kb) that are very highly expressed (≥1000 TPM) showed that the 3’ and 5’ ends of these genes were spatially separated and still had bursty expression (Leidescher et al., 2020). When the genes were inhibited, the 3’-5’ sites came closer together. Polymer modelling suggest this could be due to the increased stiffness thanks to the presence of many RNA polymerases on the gene. This suggests that for these very long, and very active genes, 3’-5’ crosstalk driven polymerase recycling may not influence transcriptional bursting. The hypothesis I am testing does not
propose that 3'-5' crosstalk is the sole cause of transcriptional noise, and the results presented from Leidescher et al. (2020) may suggest that the results from my thesis are more applicable to genes of shorter length and more intermediate expression. These genes may be less stiff with fewer RNA polymerases.

A second question I aim to answer in this thesis is if the rate of abortive transcription changes along a model gene. I have the hypothesis that abortive transcription rates decrease along genes. This hypothesis is based on an evolutionary argument (discussed further in Part 2, Chapter 4), that more resources are wasted producing a truncated non-functional mRNA later in productive elongation than earlier. This may give an evolutionary pressure to reduce abortive transcription rates along a gene.

Thesis overview

In Part 1 of this thesis, I demonstrate that 3'-5' crosstalk contributes towards transcriptional noise. I collaborate on a study measuring transcriptional noise on two model genes which have a point mutation in the poly-adenylation signal to disrupt 3'-5' crosstalk (Part 1, Chapter 1). I find that this mutation causes a reduction in transcriptional noise (Cavallaro et al., 2021). I follow this by taking a bioinformatics approach, utilising previously published sequencing data to measure 3'-5' crosstalk and transcriptional noise on a genome wide scale (Part 1, Chapter 2). Here, I also find that 3'-5' crosstalk contributes towards transcriptional noise (Cavallaro et al., 2021).

In Part 2, I develop a novel technique to label, hold, and track Pol2 from a specific genic location (Part 2, Chapter 1). I develop a simplified model of transcription, which I used to show that this technique can be used to measure recycling (Part 2, Chapter 2). I use this novel technique on the perturbated cell lines containing integrated HBB with and without the poly-adenylation signal point mutation to test if the proposed mechanism of polymerase recycling contributes to this transcriptional noise (Part 2, Chapter 3). Finally, I use this technique to test the relative rates of abortive transcription along a model gene (Part 2, Chapter 4).
Materials and Methods

Cell maintenance

Cells were certified mycoplasma free (PCR test) and maintained in Dulbecco’s Modified Eagle Medium (DMEM) + 10% FBS at 37°C and 5% CO₂. They were split with trypsin/versine and kept for low passage numbers (<15). The cells were stored by freezing in cell culture medium + 5% DMSO in either -80°C for short term storage or liquid nitrogen for long term storage.

Western Blotting

The reagents and their composition/source for Western blotting is in table 1.

Cells were grown to approximately 70% confluency on a 6 well dish and lysed in RIPA buffer (130 µl) and protease inhibitors (1.3 µl) for 5 mins with swirling on ice. The cells were scraped and centrifuged at 14000xg to collect the supernatant. The supernatant was heated with Laemmli buffer (4x) for 5mins at 95°C and 20 µl was loaded onto a SDS Page gel (4-12%). A biotinylated molecular weight ladder and/or a broad protein ladder was added, and the gel was run for 100 mins at 100 volts. Semi-dry transfer to a nitrocellulose membrane was performed. The membrane was cut in two to simultaneously stain for GAPDH and biotinylated Pol2.

The top half was incubated in blocking buffer (filtered 5% BSA in TBST) for 1 hour at room temperature and incubated with streptavidin:HRP overnight at 4 degrees (1:1000) or an anti-biotin antibody (1:1000) both diluted in blocking buffer. The membrane was then washed in TBST for 5 mins 3 times and acquired with chemiluminescence.

The bottom half was blocked for 1 hour at room temperature (filtered 5% BSA in TBST) and incubated with GAPDH antibody in blocking buffer (1:1000) overnight at 4 degrees. The membrane was then washed in TBST for 5 mins 3 times and incubated
with Anti-mouse kappa binding protein (1:10000) in blocking buffer for 1 hour at room temperature. The membrane was then washed in TBST for 5 mins 3 times and acquired with chemiluminescence.

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Composition/Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>RIPA buffer</td>
<td><a href="https://www.thermofisher.com/order/catalog/product/89900#89900">https://www.thermofisher.com/order/catalog/product/89900#89900</a></td>
</tr>
<tr>
<td>Protease inhibitors</td>
<td><a href="https://www.thermofisher.com/order/catalog/product/78429#78429">https://www.thermofisher.com/order/catalog/product/78429#78429</a></td>
</tr>
<tr>
<td>Precast gel</td>
<td><a href="https://www.abcam.com/teo-tricine-precast-gels-runbluetrade-4-12-12-well-8x10cm-ab139596.html">https://www.abcam.com/teo-tricine-precast-gels-runbluetrade-4-12-12-well-8x10cm-ab139596.html</a></td>
</tr>
<tr>
<td>BSA</td>
<td><a href="https://uk.vwr.com/store/product/7996265/bovine-serum-albumin-bsa">https://uk.vwr.com/store/product/7996265/bovine-serum-albumin-bsa</a></td>
</tr>
<tr>
<td>GAPDH antibody</td>
<td><a href="https://www.ptglab.com/products/GAPDH-antibody-60004-1-lg.htm">https://www.ptglab.com/products/GAPDH-antibody-60004-1-lg.htm</a></td>
</tr>
<tr>
<td>Western Blotting Substrate</td>
<td><a href="https://www.thermofisher.com/order/catalog/product/32106">https://www.thermofisher.com/order/catalog/product/32106</a></td>
</tr>
</tbody>
</table>

Table 1. Reagents for Western blot

ChIP-Seq experiments

The reagents and their composition/source for ChIP-Seq experiments is in table 2.
Protease inhibitors are added to PBS, and buffers A2, A-D (Table 2) and tris, by adding a protease inhibitor tablet to 50 ml of buffer. All buffers are kept on ice.

**HEK293 cell preparation – short time periods (up to 1 hour)**

Cells were grown to approximately 60% confluence on a 100mm plate, transfected with 1.48 μg of both plasmids (Avi-RPB1, TetR-BirA) in turbofect (8.80 μl) and 586 μl FBS free DMEM and left to grow overnight. They were then washed with PBS, induced in tetracycline diluted in PBS (250 ng/ml) for the time points at 37°C, fixed for 10 minutes in 4% formaldehyde in PBS and quenched in glycine (1ml of 2.5M) for five minutes. The cells were washed in cold PBS with protease inhibitors, collected by scraping and pelleted by centrifugation (1000xg, 5 mins). They were then flash frozen in a dry ice/IMS slurry and stored in -80°C.

**HEK293 cell preparation – long time periods**

1 million cells were seeded per 100mm dish for each sample and left to grow overnight. They were then forward transfected with 1.48 μg, 148 ng or 14.8 ng of Avi-RPB1 and TetR-BirA. An empty Px458 plasmid was used to keep total DNA mass transfected constant (1.48 μg). Plasmid DNA was diluted in 8.8 μl turbofect and 586 μl FBS free DMEM and added to the cells. The next day, each dish was split into 2 100mm dishes (i.e., 2.5ml in 10ml) and 24 hours later were induced/not induced with tetracycline (250ng/ml in DMEM) for 24 hours. The cells were washed with PBS, fixed for 10 minutes in 4% formaldehyde in PBS and quenched in glycine (1ml of 2.5M) for five minutes. The cells were washed in cold PBS with protease inhibitors, collected by scraping and pelleted by centrifugation (1000xg, 5 mins). They were then flash frozen in a dry ice/IMS slurry and stored in -80°C.

**HeLa cells experiment**

1 million cells were seeded per 100mm dish for each sample and left to grow overnight. They were then forward transfected with 1.48 μg of Avi-RPB1, TetR-BirA
and one of: TetO2-intron, TetO2-exon, TetO2-5’, TetO2-3’ in 8.8 μl viafect and 586 μl FBS free DMEM. The next day, they were 50% split into 2 100mm dishes (i.e. 2.5ml in 10ml) and 24 hours later were induced/not induced with tetracycline (250ng/ml in DMEM) for 24 hours. The cells were washed with PBS, fixed for 10 minutes in 4% formaldehyde in PBS and quenched in glycine (1ml of 2.5M) for five minutes. The cells were washed in cold PBS with protease inhibitors, collected by scraping and pelleted by centrifugation (1000xg, 5 mins). They were then flash frozen in a dry ice/IMS slurry and stored in -80°C.

**ChIP pulldown**

Streptavidin beads were prepared by washing 50 μl beads per sample with A2 three times (1 ml), and incubated overnight in 300 μl blocking buffer (A2 + 0.5% BSA). The cells were collected, thawed on ice, and lysed in buffer A2 with protease inhibitors (300 μl) and sonicated with a Bioruptor® shearing device on ice (30s on/30s off, high power, 30mins). The samples were centrifuged to remove debris (max speed 30 mins) and incubated overnight at 4°C with washed streptavidin beads in blocking buffer (A2 + 0.5% BSA).

ChIPed samples were then washed with wash buffers A-D, then Tris (1mm pH 8.0) then 3 times in RIPA (all 1ml) and eluted in EB containing proteinase K (0.4mg/ml) and RNAse A (0.2mg/ml) overnight at 65°C.

The DNA was purified with a DNA clean-up kit and the NEB next ultra2 protocol was followed for library preparation, with 13 PCR cycles. The library was sequenced on the Illumina platform.

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Composition/Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turbofect transfection reagent</td>
<td><a href="https://www.thermofisher.com/order/catalog/product/R0532">https://www.thermofisher.com/order/catalog/product/R0532</a></td>
</tr>
<tr>
<td>Reagent</td>
<td>Link</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Viafect transfection reagent</td>
<td><a href="https://www.promega.co.uk/products/luciferase-assays/transfection-reagents/viafect-transfection-reagent/">https://www.promega.co.uk/products/luciferase-assays/transfection-reagents/viafect-transfection-reagent/</a></td>
</tr>
<tr>
<td>Streptavidin beads</td>
<td><a href="https://www.thermofisher.com/order/catalog/product/1205D">https://www.thermofisher.com/order/catalog/product/1205D</a></td>
</tr>
<tr>
<td>DNA clean-up kit</td>
<td><a href="https://international.neb.com/products/t1030-monarch-pcr-dna-clean-up-kit-5-ug#Product%20Information">https://international.neb.com/products/t1030-monarch-pcr-dna-clean-up-kit-5-ug#Product%20Information</a></td>
</tr>
<tr>
<td>A2</td>
<td>15 mM HEPES, pH 7.5, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.5% N-lauroylsarcosine, 0.1% sodium deoxycholate, 0.1% SDS</td>
</tr>
<tr>
<td>Wash A</td>
<td>10 mM Tris pH 8.0, 1 mM EDTA, 0.1% Triton X-100, 0.2% SDS</td>
</tr>
<tr>
<td>Wash B</td>
<td>20 mM Tris pH 8.0, 150 mM NaCl, 5 mM EDTA, 5.2% sucrose, 1% Triton X-100, 0.2% SDS</td>
</tr>
<tr>
<td>Wash C</td>
<td>5 mM Tris pH 8.0, 25 mM HEPES, 250 mM NaCl, 0.5 mM EDTA, 0.5% Triton X-100, 0.05% sodium deoxycholate</td>
</tr>
<tr>
<td>Wash D</td>
<td>10 mM Tris pH 8.0, 250 mM LiCl, 10 mM EDTA, 0.5% IGEPAL CA-630, 0.5% sodium deoxycholate</td>
</tr>
<tr>
<td>Tris</td>
<td>10 mM pH 8.0</td>
</tr>
<tr>
<td>RIPA</td>
<td>50 mM HEPES, 500 mM LiCl, 1 mM EDTA, 0.7% sodium deoxycholate, 1% IGEPAL CA-630</td>
</tr>
<tr>
<td>EB</td>
<td>25 mM Tris pH 8.0/ 5 mM EDTA, 0.5% SDS</td>
</tr>
<tr>
<td>BSA</td>
<td><a href="https://uk.vwr.com/store/product/7996265/bovine-serum-albumin-bsa">https://uk.vwr.com/store/product/7996265/bovine-serum-albumin-bsa</a></td>
</tr>
</tbody>
</table>

Table 2. Reagents for ChIP-Seq experiments
ELISA

The reagents and their composition/source for ELISA experiments is in table 3.

1 million cells were seeded in a 100mm dish and transfected the next day with 1.48 μg of both plasmids (Avi-RPB1, TetR-BirA) in turbofect (8.80 μl) and left growing overnight. They were split 25% and then left for 2 days to match the starting conditions of the ChIP-qPCR experiments. The dishes were washed in ice cold PBS, lysed in RIPA buffer (800 μl) and protease inhibitors (8 μl) for 5 mins with swirling on ice. The total protein concentration was then measured with a BCA assay according to the manufacturer’s instructions.

60 μg total protein for each sample was loaded and a dilution series of standards was prepared. The manufacture’s recommended protocol was followed and the resulting optical density was measured with a plate reader. The standard curve was created using linear regression in R (R Core Team, 2020) and used to calculate the starting concentrations of Pol2.

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Composition/Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turbofect</td>
<td><a href="https://www.thermofisher.com/order/catalog/product/R0532">https://www.thermofisher.com/order/catalog/product/R0532</a></td>
</tr>
<tr>
<td>RIPA</td>
<td><a href="https://www.thermofisher.com/order/catalog/product/89900">https://www.thermofisher.com/order/catalog/product/89900</a></td>
</tr>
<tr>
<td>Protease inhibitor</td>
<td><a href="https://www.thermofisher.com/order/catalog/product/78429">https://www.thermofisher.com/order/catalog/product/78429</a></td>
</tr>
<tr>
<td>BCA assay</td>
<td><a href="https://www.thermofisher.com/order/catalog/product/23225">https://www.thermofisher.com/order/catalog/product/23225</a></td>
</tr>
</tbody>
</table>

*Table 3. Reagents for ELISA*
Proteomics

The reagents and their composition/source for the proteomics mass spectrometry experiments is in table 4.

1 million cells were seeded in a 100mm dish and transfected the next day with 1.476 μg of both plasmids (Avi-RPB1, TetR-BirA) in turbofect (8.80 μl) and left growing overnight. They were split 25% and then left for 2 days to match the starting conditions of the ChIP-qPCR experiments. The dishes were washed in ice cold PBS, collected by scraping, and pelleted by centrifugation at 1000xg for 5 mins. They were either flash frozen and stored in -80°C or the experiment continued.

Streptavidin beads were prepared by washing 50 μl beads per sample with A2 three times (1 ml) and incubated overnight in 300 μl blocking buffer (A2 + 0.5% BSA). The cells were thawed on ice and lysed in buffer A2 with protease inhibitors (300 μl) and sonicated with a Bioruptor® shearing device on ice (30s on/30s off, high power, 30mins). The sample was spun to remove debris (max speed, 30 mins) and was incubated overnight with washed streptavidin beads in blocking buffer (A2 + 0.5% BSA) at 4°C.

The samples were washed with buffers A-D, Tris (1mM, pH 8.0), RIPA and Tris (1mM, pH 8.0) 3 times.

The samples were washed with Ammonium bicarbonate (ABC, 50mM) and reduced/alkylated with tris(2-carboxyethyl)phosphine (TCEP, 10mM) and 2-Chloroacetamide (CAA 40mM) in ABC for 5 mins at 70°C. They were then digested with 0.5 μg trypsin overnight at 37 °C and acidified with 5μL Trifluoroacetic acid (TFA, 1%).

The samples were purified with StageTip following the protocol of Rappsilber et al. (2007). Briefly, 2 layers of a Octadecylsilane (C18) membrane were put into a 200μL pipette tip and conditioned with methanol, followed by Acetonitrile (ACN) and then
2%ACN + 0.1%TFA. The samples were loaded, washed with Ethyl Acetate + 1%TFA, followed by 2%ACN+0.1%TFA. They were eluted in 60% ACN + 1%TFA, dried in a speedVac and resuspended in 2%ACN + 0.1%TFA which was then used for mass spectrometry.

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Composition/Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turbofect transfection reagent</td>
<td><a href="https://www.thermofisher.com/order/catalog/product/R05">https://www.thermofisher.com/order/catalog/product/R05</a> 32</td>
</tr>
<tr>
<td>Streptavidin beads</td>
<td><a href="https://www.thermofisher.com/order/catalog/product/112">https://www.thermofisher.com/order/catalog/product/112</a> 05D</td>
</tr>
<tr>
<td>A2</td>
<td>15 mM HEPES, pH 7.5, 140 mM NaCl, 1mM EDTA, 1% Triton X-100, 0.5% N-lauroylsarcosine, 0.1% sodium deoxycholate, 0.1% SDS</td>
</tr>
<tr>
<td>Wash A</td>
<td>10 mM Tris pH 8.0, 1mM EDTA, 0.1% Triton X-100, 0.2%SDS</td>
</tr>
<tr>
<td>Wash B</td>
<td>20 mM Tris pH 8.0, 150 mM NaCl, 5 mM EDTA, 5.2% sucrose, 1% Triton X-100, 0.2% SDS</td>
</tr>
<tr>
<td>Wash C</td>
<td>5 mM Tris pH 8.0, 25 mM HEPES, 250 mM NaCl, 0.5 mM EDTA, 0.5% Triton X-100, 0.05% sodium deoxycholate</td>
</tr>
<tr>
<td>Wash D</td>
<td>10 mM Tris pH 8.0, 250 mM LiCl, 10 mM EDTA, 0.5% IGEPAL CA-630, 0.5% sodium deoxycholate</td>
</tr>
<tr>
<td>Tris</td>
<td>10 mM pH 8.0</td>
</tr>
<tr>
<td>RIPA</td>
<td>50 mM HEPES, 500 mM LiCl, 1 mM EDTA, 0.7% sodium deoxycholate, 1% IGEPAL CA-630</td>
</tr>
<tr>
<td>BSA</td>
<td><a href="https://uk.vwr.com/store/product/7996265/bovine-serum-albumin-bsa">https://uk.vwr.com/store/product/7996265/bovine-serum-albumin-bsa</a></td>
</tr>
<tr>
<td>ABC</td>
<td>Ammonium bicarbonate 50 mM</td>
</tr>
<tr>
<td>TCEP</td>
<td>tris(2-carboxyethyl)phosphine 10mM</td>
</tr>
</tbody>
</table>
Gibson assembly

The reagents and their composition/source for Gibson assembly is in table 5.

The sequence for the gene HBB was downloaded from NCBI (Assembly GRCh38.p13) and manually annotated. gblocks were purchased from IDT (covering 1 – 5’UTR, exon 1, intron 1 exon 2, 2 – intron 2, 3 – exon 3 3’UTR along with various configurations containing TetO2 in different locations). Overlapping regions were generated with PCR, and the primers were designed in snapgene with the Gibson assembly tool.

Gibson assembly followed the manufacturer’s (NEB) protocol. Simply, my vector cFLAG PCDNA3 (2 µg) (Addgene plasmid:20011) was linearised with HINDIII-HF (2 µl) and EcoR1 HF (2 µl) in NEBuffer 3.1 (5 µl) and diluted in water to a 50 µl reaction volume. Successful linearisation was tested with gel electrophoresis.

Fragments containing overlapping regions were added at a 3:1 molar ratio (fragments:vector), Gibson master mix was added and incubated for 1 hour at 50°C. The product was transformed into Sig10 cells and colony PCR was used to screen for successful assembly. Plasmids were verified with Sanger sequencing.

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Composition/Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCR HS VeriFi Master Mix</td>
<td><a href="https://pcrbio.com/products/pcr/pcrbio-hs-verifi-polymerase-mixes/">https://pcrbio.com/products/pcr/pcrbio-hs-verifi-polymerase-mixes/</a></td>
</tr>
<tr>
<td>cFLAG PCDNA3</td>
<td><a href="https://www.addgene.org/20011/">https://www.addgene.org/20011/</a></td>
</tr>
<tr>
<td>HIND-III HF</td>
<td><a href="https://international.neb.com/products/r3104-hindiii-hf#Product%20Information">https://international.neb.com/products/r3104-hindiii-hf#Product%20Information</a></td>
</tr>
<tr>
<td>Reagent</td>
<td>URL</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>EcoR1 HF</td>
<td><a href="https://international.neb.com/products/r3101-ecori-hf#Product%20Information">https://international.neb.com/products/r3101-ecori-hf#Product%20Information</a></td>
</tr>
<tr>
<td>NEBuffer 3.1</td>
<td><a href="https://international.neb.com/products/b7203-nebuffer-3-1#Product%20Information">https://international.neb.com/products/b7203-nebuffer-3-1#Product%20Information</a></td>
</tr>
<tr>
<td>Gibson Assembly Master Mix</td>
<td><a href="https://international.neb.com/products/e2611-gibson-assembly-master-mix#Product%20Information">https://international.neb.com/products/e2611-gibson-assembly-master-mix#Product%20Information</a></td>
</tr>
<tr>
<td>Sig10 competent cells</td>
<td><a href="https://www.sigmaaldrich.com/GB/en/product/sigma/cmc0001">https://www.sigmaaldrich.com/GB/en/product/sigma/cmc0001</a></td>
</tr>
</tbody>
</table>

Table 5. Reagents for Gibson assembly

ChIP-qPCR experiment

The reagents and their composition/source for the ChIP-qPCR experiments is in table 6.

1 million cells were seeded in a 100mm dish and transfected the next day with 1.476 μg of both plasmids (Avi-RPB1, TetR-BirA) in turfobfect (8.80 μl) and left to grow overnight. They were fully split into 4 100mm dishes and the day after they were induced in tetracycline (250ng/ml) for the time points. This was done for all time points to finish at the same time. The dishes were washed in PBS, fixed for 10 minutes in 4% formaldehyde in PBS and quenched in glycine (1ml of 2.5M) for five minutes. The cells were washed in cold PBS with protease inhibitors and collected by scraping. They were pelleted by centrifugation at 5 mins 1000 x g, flash frozen in a dry ice/IMS slurry and stored in -80°C.

Streptavidin beads were prepared by washing 50 μl streptavidin M280 beads per sample with A2 three times (1 ml) and incubating overnight in 300 μl blocking buffer (A2 + 0.5% BSA). Pol2 beads were prepared by washing 50 μl Protein G beads per sample with A2 three times (1 ml) and incubating overnight in 5 μg antibody in 300 μl blocking buffer (A2 + 0.5% BSA).

The cells were collected, thawed on ice and lysed in buffer A2 with protease inhibitors (300 μl) and sonicated with a Bioruptor® shearing device on ice (30s on/30s
off, high power, 30mins). The cells were centrifuged to remove debris (max speed 30 mins) and incubated overnight with antibody bound or streptavidin beads in blocking buffer (A2 + 0.5% BSA) for Pol2 chip and biotin Pol2 chip respectively. The ChIPed samples were washed with buffers A-D, Tris (1mM, pH 8.0), then with RIPA three times. The samples were then eluted in EB containing proteinase K (0.4mg/ml) and RNase A (0.2mg/ml). A DNA clean-up kit was used to purify the DNA. 100pg of purified DNA was used for qPCR, along with 500nM forward and reverse primers and 5μl PowerUp™ SYBR™ Green Master Mix in a 10μl reaction volume. qPCR was conducted using the Rotor-Gene Q system and analysed with the qpcR package v1.4-1 (Spiess, 2018).

<table>
<thead>
<tr>
<th>Reagent</th>
<th>Composition/Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turbofect transfection reagent</td>
<td><a href="https://www.thermofisher.com/order/catalog/product/R0532">https://www.thermofisher.com/order/catalog/product/R0532</a></td>
</tr>
<tr>
<td>Pol2 Antibody</td>
<td><a href="https://www.scbt.com/p/pol-ii-antibody-ctd4h8">https://www.scbt.com/p/pol-ii-antibody-ctd4h8</a></td>
</tr>
<tr>
<td>Streptavidin beads</td>
<td><a href="https://www.thermofisher.com/order/catalog/product/11205D">https://www.thermofisher.com/order/catalog/product/11205D</a></td>
</tr>
<tr>
<td>Protein G beads</td>
<td><a href="https://www.thermofisher.com/order/catalog/product/10003D">https://www.thermofisher.com/order/catalog/product/10003D</a></td>
</tr>
<tr>
<td>DNA clean-up kit</td>
<td><a href="https://international.neb.com/products/t1030-monarch-pcr-dna-cleanup-kit-5-ug#Product%20Information">https://international.neb.com/products/t1030-monarch-pcr-dna-cleanup-kit-5-ug#Product%20Information</a></td>
</tr>
<tr>
<td>A2</td>
<td>15 mM HEPES, pH 7.5, 140 mM NaCl, 1mM EDTA, 1% Triton X-100, 0.5% N-lauroylsarcosine, 0.1% sodium deoxycholate, 0.1% SDS</td>
</tr>
<tr>
<td>Wash A</td>
<td>10 mM Tris pH 8.0, 1mM EDTA, 0.1% Triton X-100, 0.2% SDS</td>
</tr>
<tr>
<td>Wash B</td>
<td>20 mM Tris pH 8.0, 150 mM NaCl, 5 mM EDTA, 5.2% sucrose, 1% Triton X-100, 0.2% SDS</td>
</tr>
<tr>
<td>Wash C</td>
<td>5 mM Tris pH 8.0, 25 mM HEPES, 250 mM NaCl, 0.5 mM EDTA, 0.05% sodium deoxycholate</td>
</tr>
<tr>
<td>Wash D</td>
<td>10 mM Tris pH 8.0, 250 mM LiCl, 10 mM EDTA, 0.5% IGEPAL CA-630, 0.5% sodium deoxycholate</td>
</tr>
<tr>
<td>RIPA</td>
<td>50 mM HEPES, 500 mM LiCl, 1 mM EDTA, 0.7% sodium deoxycholate, 1% IGEPAL CA-630</td>
</tr>
<tr>
<td>EB</td>
<td>25 mM Tris pH 8.0/ 5 mM EDTA, 0.5% SDS</td>
</tr>
<tr>
<td>Proteinase K</td>
<td><a href="https://www.thermofisher.com/order/catalog/product/E00491">https://www.thermofisher.com/order/catalog/product/E00491</a></td>
</tr>
<tr>
<td>Reagent</td>
<td>Link</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>BSA</td>
<td><a href="https://uk.vwr.com/store/product/7996265/bovine-serum-albumin-bsa">https://uk.vwr.com/store/product/7996265/bovine-serum-albumin-bsa</a></td>
</tr>
<tr>
<td>RNAse A</td>
<td><a href="https://www.thermofisher.com/order/catalog/product/E.N0531">https://www.thermofisher.com/order/catalog/product/E.N0531</a></td>
</tr>
<tr>
<td>QPCR master mix</td>
<td><a href="https://www.thermofisher.com/order/catalog/product/A25741">https://www.thermofisher.com/order/catalog/product/A25741</a></td>
</tr>
<tr>
<td>GAPDH promoter forward primer</td>
<td>CTCCTGTTGACAGTCAGCC</td>
</tr>
<tr>
<td>GAPDH promoter reverse primer</td>
<td>TCAGGCGCTCCCCTAGCC</td>
</tr>
<tr>
<td>HBB promoter forward primer</td>
<td>ATCGTCGACGAGCTCGTTTAG</td>
</tr>
<tr>
<td>HBB promoter reverse primer</td>
<td>CTAGTGAACACAGTTGTGTCAGAAG</td>
</tr>
</tbody>
</table>

*Table 6. Reagents for ChIP-qPCR*

**Custom R scripts**

Custom R (R Core Team, 2020) scripts were used for much of my analysis. Data manipulation was performed with base R or the packages from tidyverse (Wickham et al., 2019). Statistics were calculated using the functions from base R, or specialised packages, referred to in text. ggplot2 was used to plot much of my results (Wickham et al., 2019).

**Influenza infected**

UMI count data from a scRNA seq experiment of influenza infected cells (A564) was downloaded from the Gene Expression Omnibus (GSE108041). This data was filtered using Seurat (Butler et al., 2018) to select truly infected cells using the same process as Russell et al. (2018). A cell was considered infected if it contained more than X% UMIIs from the influenza virus (where X is 0.02 for 6 hours and 8 hours replicate 1, 0.043 for 8 hours replicate 2 and 0.14 for ten hours). PYMC (Patil et al., 2010) was used to fit a negative binomial distribution to the filtered scRNA-seq UMI counts using the approach of Cavallaro et al. (2021) to estimate transcriptional parameters. Genes were filtered to only include transcriptionally active genes (μ > 0.05, CV^2 <3000). Custom R scripts were then used to estimate the noise between infected groups.
Antisense transcription

Raw reads from a stranded RNA-seq dataset for the HBB WT and pA- lines after tetracycline induction (Cavallaro et al., 2021) were downloaded (SRA identifier: SRR8394947 and SRR8394953). These were aligned to the modified reference genome. LiBiNorm (Dyer et al., 2019) was used to count reads aligning to HBB, corresponding to sense and antisense transcription.

ChIA-PET - 3’-5’ crosstalk

K562 CHIA-PET raw sequencing reads from different cell lines (SRA identifiers K564: SRR372747, SRR372748, SRR1514656. MCF7: SRR372741) were downloaded, trimmed with Trim_galore (Krueger, 2021) and analysed with CHIA-PET2 using default parameters (Li et al., 2017). This generated Hi-C style interaction matrices of varying resolutions. Reads corresponding to bins containing the start and end of a gene (O’Leary et al., 2016) were extracted. This was then normalised by gene length by dividing the number of reads by the average read number from 10,000 random genomic intervals of the same length from the same chromosome. This was transformed by arcsinh √(x + 0.5), resulting in a variance-stable 3’-5’ interaction score (Bartlett, 1947). Custom R scripts were used to compare this to transcriptional parameters inferred from scRNA-Seq.

scRNA-seq analysis - 3’-5’ crosstalk

UMI count data from a scRNA-Seq dataset of K562 cells was downloaded from the Gene Expression Omnibus (GSE124682). Massimo fitted this UMI count data to a negative binomial distribution using PYMC (Patil et al., 2010) allowing parameters of transcription to be estimated.

Gene Ontology – 3’-5’ crosstalk

A gene was defined as looped if it is in the top 10% of 3’-5’ interaction scores. The package enrichR (Kuleshov et al., 2016) was used discover enriched functions of looped genes, using the database GO Molecular Function 2015. A significance
threshold of adjusted P value < 0.05 was used. This was repeated for 4 ChIA-PET datasets analysed. Custom R scripts were used to compare these functions to inferred transcriptional parameters.

Network analysis – 3’-5’ crosstalk

Hi-C style contact matrices at 2 kb resolution were treated as a network, with each node representing a 2 kb bin of the genome and links defined as ChIA-PET interactions between the nodes. Nodes with no links were excluded from this analysis, and custom R scripts were used to calculate the degree of each node (the number of other nodes each node is linked to). The PoweRlaw R package (Gillespie, 2015) was used to fit a variety of heavy tailed distributions to the degree distribution, and to calculate statistics of best fitting distributions. Custom R scripts were used to plot this data, and determine if nodes overlapped with genes (O’Leary et al., 2016) or enhancers (Gao and Qian, 2020).

Pol2 occupancy – 3’-5’ crosstalk

RPKM normalised coverage files of total Pol2 ChIP-Seq and ChIP-Exo experiments were downloaded from the Gene Expression Omnibus (Accession: GSE108323) (McHaourab et al., 2018). Average read coverage at the 5’ end (TSS +/- 250 bp) and the 3’ end (TES +/- 250 bp) at all genes were calculated using the program bwaverageoverbed (Kuhn et al., 2013). Custom R scripts were used to compare this to transcriptional parameters.

ChIP-Seq HEK293 Cell lines

The hg38 reference genome (Schneider et al., 2017) was modified by adding the insert into its correct position (Cavallaro et al., 2021), using reform (Khalfan et al., 2018). The transfected plasmids were treated as separate chromosomes and added to the reference genome. Genomic HBB, POL2RA and areas with unreasonably high coverage were suppressed with bedtools maskfasta (Quinlan and Hall, 2010). The associated annotation file was also modified.
The raw reads were quality checked with fastqc (Andrews, 2010), quality and adaptor trimmed with trim_galore (Krueger, 2021). They were aligned to the modified reference genome using bowtie2 (Langmead and Salzberg, 2013) with the --very-sensitive parameter. The resulting sequence alignment map was converted into a binary compressed version and indexed with Samtools (Li et al., 2009). IGV was used to visualise the alignment (Robinson et al., 2011).

Deeptools (Ramírez et al., 2016) was used to create CPM normalised coverage files, Log2 fold ratio coverage files as well as metagene plots. MACS2 (Zhang et al., 2008) was used to call peaks under default parameters and these peaks were analysed with custom R scripts, using the packages ChIPpeakAnno and genomation (Zhu et al., 2010; Akalin et al., 2015). Coverage at promoter proximal region was calculated as TSS +/- 250 bp using the program bwaverageoverbed (Kuhn et al., 2013).

Model

A system of ordinary differential equations (ODE’s) was constructed to describe this model. These were implemented using custom R scripts, using the package DeSolve (Karline Soetaert, Thomas Petzoldt, 2010).

Proteomics

Database searching: All MS/MS samples were analysed using MaxQuant version 2.0.3.0 (Cox and Mann, 2008). MaxQuant was set up to search a custom database containing the fusion protein Avi-RPB1 as well as the Universal protein resource knowledge base assuming the digestion enzyme strict-trypsin. MaxQuant was searched with a fragment ion mass tolerance of 0.50 Da and a parent ion tolerance of 20 PPM. Carbamidomethyl of cysteine was specified in MaxQuant as a fixed modification. Oxidation of methionine and acetyl of the n-terminus were specified in MaxQuant as variable modifications.
Quantitative analysis: Scaffold version Scaffold_5.1.0 (Searle, 2010) was used to validate MS/MS based peptide and protein identifications. Peptide identifications were accepted if they could be established at greater than 95.0% probability by the Percolator posterior error probability calculation (Käll et al., 2008). Protein identifications were accepted if they could be established at greater than 95.0% probability and contained at least 2 identified peptides. Protein probabilities were assigned by the Protein Prophet algorithm (Nesvizhskii et al., 2003). Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony. Proteins sharing significant peptide evidence were grouped into clusters.

Normalization was performed iteratively (across samples) on intensities (Hill et al., 2008; Oberg et al., 2008). Means were used for averaging. Spectra data were log-transformed and weighted by an adaptive intensity weighting algorithm. Of 1,823 spectra in the experiment at the given thresholds, 488 (27%) were included in quantitation. Differentially expressed proteins were determined by applying t-test with unadjusted significance level p < 0.05.

ChIP-Seq HeLa Cell lines

The hg38 reference genome (Schneider et al., 2017) was modified by adding the transfected plasmids, which were each treated as a separate chromosome. This included the plasmids encoding TetR-BirA, Avi-RPB1 and HBB with the Tet operon in the correct location. The raw reads were quality checked with fastqc (Andrews, 2010), quality and adaptor trimmed with trim_galore (Krueger, 2021). They were aligned to the modified reference genome using bowtie2 (Langmead and Salzberg, 2013) with the --very-sensitive parameter. The resulting sequence alignment map was converted into a binary compressed version and indexed with Samtools (Li et al., 2009). Deeptools (Ramírez et al., 2016) was used to create Log₂ fold ratio coverage files as well as metagene plots. The bedtools program bwaverageoverbed (Kuhn et al., 2013) was used to calculate the average log₂ fold ratio at the promoter proximal region (TSS +/- 250 bp) for genes. Custom R scripts were used to analyse this ratio.
Part 1: 3’-5’ crosstalk contributes towards transcriptional noise

Chapter 1: An approach with two model genes

Introduction

In this section of the thesis, I will describe the experiments I undertook in collaboration with other members of our laboratory to demonstrate that 3’-5’ crosstalk contributes towards transcriptional noise.

The first experiment I will mention is an experiment interrogating how transcriptional noise changes when 3’-5’ crosstalk is interrupted. This crosstalk was interrupted through mutation of the poly-adenylation signal in two model genes in HEK293 T-Rex cell lines. These cell lines stably express the Tet repressor protein.

The first pair of cell lines we used contain an ectopic insertion of HBB under the control of a tetracycline responsive promoter (referred to as HBB WT), and another cell line contains the same insertion with a point mutation at the SV40 poly-Adenylation signal; A₂TA₃ is mutated to A₂GA₃ (this cell line is referred to as HBB pA⁻).

The second pair of cell lines are also HEK293 T-Rex lines which contain an insertion of HIV env, also under the control of a tetracycline responsive promoter (HIV env WT). The other cell line contains the same insertion with a point mutation in the BGH poly-adenylation signal; A₂TA₃ is mutated to A₂GA₃ (HIV env pA⁻). The inserted genes from both pairs of cell lines are induced by the addition of tetracycline.

Previous work has shown that mutation of these poly-adenylation signals reduce Pol2 at the 5’ end of a gene (Mapendano et al., 2010). This suggests there is crosstalk between the 5’ and 3’ end of a gene. The authors also demonstrated reduced steady
state levels of mRNA for these genes, suggesting that this crosstalk is transcriptionally relevant.

Indeed, a mechanism for this 3'-5' crosstalk has been measured directly in HBB for both the WT and pA- cell lines. A 3C assay has shown that the 5' and the 3' end of HBB are in close three-dimensional proximity in the WT line, but not in the pA- line (Tan-Wong et al., 2012). However, previous work in our group has been unable to reproduce this. This suggests that the crosstalk between the 3'-5' ends of the gene is caused by physical proximity and provides further evidence of crosstalk being disrupted by mutating the poly-adenylation signal.

Work in our laboratory showed that when the poly-adenylation signal was mutated, the mean RNA expression of HBB after tetracycline induction was decreased (Cavallaro et al., 2021). Briefly, this was measured through short fluorescently labelled probes constructed against HBB and HIV env which were hybridised to HBB and HIV env mRNA, and the fluorescence was measured in the flow cytometer, a technique known as FLOW-FISH (Tiberi et al., 2018). The fluorescence for the WT lines was on average higher than the pA- lines for the same levels of tetracycline induction, which shows the poly-adenylation signal point mutation reduced mean HBB expression (Figure 3). These experiments were conducted by Mark Walsh. This matches the results from Mapendano et al. (2010), showing that the disruption of crosstalk is transcriptionally relevant.
Figure 3. Taken from (Cavallaro et al., 2021). Image made by Mark Walsh. Kernel density estimates of fluorescence for four replicates of FLOW-FISH experiments at different tetracycline induction concentrations (left) for two cell lines HBB and HIV-env. Each cell line had two configurations, the Wild Type gene, or a mutation in the poly-adenylation signal. Both the mean value at each tetracycline induction concentration and the width of the distributions is higher in the WT line compared to the mutant. Control shows cellular auto-fluorescence when the gene is deleted.

We wanted to be able to use these distributions of fluorescence from flow cytometry to measure transcriptional parameters. To achieve this, Massimo Cavallaro fitted these distributions to three models: the one-state model (Model A, Equation 1 A, Figure 1 A) described earlier – giving rise to a Poisson distribution of mRNA (Figure 1C), the two-state model (Model B, Equation 1 A, Figure 1 B) described earlier – giving rise to a Poisson-beta mixture distribution of mRNA (Equation 1B), which is wider than the Poisson (Figure 1 C) and a simplified version of the two state model (Golding et al., 2005). In this simplified version, the gene on-to-off rate ($k_{off}$) and the transcription rate ($\alpha$) is condensed into a single parameter, burst size, which is defined as the average number of mRNAs made per burst ($\alpha/k_{off}$). This results in a negative binomial mRNA distribution. All models apply to the steady state.
To calibrate the FLOW-FISH readings for the HBB cell lines, we used RNA smFISH (Femino et al., 1998; Raj et al., 2008) to count individual RNA molecules which would appear as a diffraction limited spot. This was used to find informative priors for the mean mRNA expression of HBB.

I conducted RNA smFISH imaging of two tetracycline concentrations on the cell lines: HBB WT 5ng/ml (48 cells) and HBB pA- 80ng/ml (179 cells). Mark Walsh from our laboratory measured the other HBB samples. We analysed the images with FISH-QUANT (Mueller et al., 2013) to count exact numbers of RNA molecules for different concentrations of tetracycline induction of HBB in both the WT and pA- cell lines. Mark Walsh used Nanostring nCounter® technology in a similar way to measure the mean expression of HIV env at each concentration of tetracycline induction.

We found that the one-state model did not fit our data well, but the other two models did. Fitting these two models (two-state and simplified two-state) to the FLOW-FISH data allowed transcriptional parameters to be inferred (Figure 4).

We measured transcriptional noise by calculating the squared coefficient of variation ($CV^2$) of mRNA levels, and found that as mRNA mean expression increased, $CV^2$ decreased (Figure 4 A-C). Eventually $CV^2$ reaches a noise floor where it does not decrease further upon increased mean expression. This has been discovered in many other systems, e.g. (Dar et al., 2015; Klein et al., 2015).

For HBB, we found that, at the same expression levels, the WT cell lines have higher transcriptional noise than the pA- cell lines. This is shown by their higher squared coefficient of variation (Figure 4 A, C). We also found that the burst frequency was higher in the pA- lines, and the burst size higher in the WT cell lines (Figure 4 D). For the HIV env cell lines, the patterns are less conclusive (Figure 4 B, C (top right), E), as the points overlap at the highest mean expression levels, but at more intermediate mean expression levels, the patterns are consistent with the HBB cell lines.
Figure 4. Taken from (Cavallaro et al., 2021). Image made by Massimo Cavallaro. Noise plots of HBB (A) and HIV env (B). (C) Consensus estimates of noise plots. These plots show transcriptional noise (CV^2) plotted against mean expression (µx). For the same expression value, the WT lines have higher transcriptional noise (in HBB) and the same for HIV-env for intermediate levels of expression. Consensus estimations of transcriptional parameters through fitting to the Poisson-beta model for HBB (D) and HIV env (E). For intermediate levels of expression, the burst frequency was higher for the mutant lines, and the burst size was higher in the WT lines. The points all correspond to medians and the 90% credible intervals of the posterior distributions.

We demonstrated, through fitting models of transcription to fluorescence readings from FLOW-FISH, that disrupting 3’-5’ crosstalk on two model genes (HBB and HIV env) results in decreased mean expression. For the same mean RNA expression, disruption of 3’-5’ crosstalk also results in reduced transcriptional noise. This provides evidence for my overall hypothesis that 3’-5’ crosstalk is one of the causes of transcriptional noise.
Results

Does the termination defect in the pA- lines cause differences in transcriptional noise?

One possible criticism for this work on the inserted genes could be that along with reduced 3’-5’ crosstalk, the pA- lines also have termination defects (Cavallaro et al., 2021), also shown in Figure 6. Here, transcription continues after the end of the gene in the pA- line. It could be argued that differences in transcriptional noise and mean expression are not caused by the disrupted 3’-5’ crosstalk but are instead caused by the termination defect.

Influenza infection also causes termination defects, where transcription of infected cells can continue for tens of kilobases after the poly-adenylation signal (Bauer et al., 2018; Zhao et al., 2018). Native elongation transcript sequencing (NET-seq) also shows that infected cells do not have a difference in initiation of transcription (Bauer et al., 2018), which suggests that 3’-5’ crosstalk is not inhibited upon infection.

I therefore aimed to test this criticism of the work on inserted genes by measuring transcriptional noise in influenza infected cells. My hypothesis was, that the influenza infected cells resemble the HEK293 (HBB or env) pA- lines in one way, namely that they have termination defects (Bauer et al., 2018; Zhao et al., 2018). However the pA- lines have reduced 3’-5’ crosstalk (Cavallaro et al., 2021; Mapendano et al., 2010), whereas influenza infected cells do not (Bauer et al., 2018). This difference therefore allows me to disentangle if 3’-5’ crosstalk or the termination defect contributes to transcriptional noise.

I analysed a scRNA-Seq dataset from influenza infected cells to answer this question (Russell et al., 2018). I implemented the approach of Cavallaro et al. (2021) to fit a negative binomial distribution to the scRNA-seq UMI counts of infected and uninfected cells. This enabled me to estimate transcriptional parameters such as CV$^2$, burst frequency ($k_{on}$), burst size and mean expression ($\mu$). I filtered these to only include transcriptionally active genes ($\mu > 0.05, CV^2 < 3000$).
I wanted to see if the transcriptional noise was different in the infected vs uninfected cells. Here, I plotted the graph of CV² against mean expression on the logarithmic axis (Figure 5 A) and estimated transcriptional noise (\(v\)) as the vertical distance from the predicted curve (\(CV^2 = 1/\mu + 1/\text{mean}(k_{on})\)), using the same approach as Cavallaro et al. (2021). Mean burst frequency was calculated as the mean from every expressed gene in each sample. I then plotted a boxplot of (\(v\)) (Figure 3 B) and found in all samples, the infected cells had significantly higher transcriptional noise than the uninfected cells (\(P<2.2e^{-16}\) for all, 2 tailed t-test).

![Figure 5. (A) Plot of CV² against UMI mean expression on the logarithmic axis, with the noise trend \(CV^2=1/\text{Mean expression} + 1/\text{mean burst frequency}\) plotted (blue curve). Each gene at each time point post infection is plotted. (B) Boxplots of transcriptional noise (\(v\), vertical distance from the predicted curve) in each group. Every infected group has significantly higher transcriptional noise than the uninfected group (\(P<2.2e^{-16}\) for all, 2 tailed t-test).](image)

These infected cell lines with termination but not 3'-5' crosstalk defects have higher transcriptional noise, but the pA- cell lines with 3'-5' crosstalk and termination defects have lower transcriptional noise than lines without the point mutation on the poly-adenylation signal. This suggests that it is not termination defects which cause less noise in the pA- cell lines and provides more evidence that it is 3'-5' crosstalk which contributes to transcriptional noise rather than defects in termination.

The work presented here formed early stages of our published results (Cavallaro et al., 2021), which was then further developed by Massimo Cavallaro. In the published
work, cells from one time point post infection (6 hours) were split into two groups, infected or uninfected (based on a threshold of 0.02% of transcripts being viral). Here, we also found that infected cells have significantly higher transcriptional noise.

**Could antisense transcription confound the results?**

Another possible confounder of the results looking at the single integrated gene from Cavallaro et al. (2021) could be antisense transcription. This can originate from the 3’ end of a gene (Malik et al., 2013) and has previously been implicated in buffering noise (Huber et al., 2016). The potential confounder could be that if antisense transcription occurs in the poly-adenylation mutant line, but not (or to a lesser extent) the WT line, then this could explain the reduced noise of the HBB gene in the pA- cell line.

I visualised RNA-seq data from the HBB cell lines (WT and pA-) after tetracycline induction (Cavallaro et al., 2021) in IGV (Robinson et al., 2011) and coloured by ‘first of pair strand’ to investigate the effect of antisense transcription. I saw minimal antisense transcription in both cell lines (Figure 6). This suggests that differing antisense transcription does not contribute to the difference in transcriptional noise between these cell lines.

This was quantified by counting the number of reads aligning to the HBB locus for both sense and antisense transcription for these two samples. Only 2.8% of reads in the pA- sample were the result of antisense transcription (253 out of 8995 total reads), and only 3.9% for the WT sample (884 out of 22923 total reads). These are both small amounts of antisense transcription, but even if this 1.1% difference between the two samples is enough to buffer noise – it would be expected to buffer noise in the WT more than the pA- mutant, opposite to what we see (Cavallaro et al., 2021). From this result, I conclude that the potential confounder of antisense transcription does not explain the reduced noise of the HBB gene in the pA- cell line.
Figure 6. A screenshot of IGV, showing a coverage plot at the integrated HBB locus, which is marked. A subsample of 10% of aligned reads were visualized. Most reads are coloured blue at this gene, showing minimal antisense transcription, this was quantified to be 3.9% for the WT cell line and 2.8% for the pA- cell line. The pA- line also has defective termination shown by the large amounts of run on transcription after the HBB gene. The ranges of counts for the coverage track are marked in brackets.

Summary

Overall, the results presented in this chapter support the hypothesis that 3’-5’ crosstalk contributes towards transcriptional noise. We found that disrupting this crosstalk by mutating the poly-adenylation signal decreased transcriptional noise, increased burst frequency and decreased burst size – for a given mean expression.

I found that the decreased transcriptional noise from the pA- cell lines was not likely due to defective termination; as an other method of disrupting termination, influenza infection, increases transcriptional noise. This is the opposite to what we see in the pA- cell lines and further suggests that 3’-5’ crosstalk contributes towards transcriptional noise. Additionally, I found that the poly-adenylation signal point mutation does not increase antisense transcription at the HBB locus, ruling out antisense transcription possibly causing this reduction in transcriptional noise.
Chapter 2: Measuring 3’-5’ crosstalk and transcriptional noise on a genome wide scale

Introduction

The previous results have shown, through an in-depth study on two model genes, that 3’-5’ crosstalk contributes towards increased mean mRNA expression and increased transcriptional noise. When this crosstalk is disrupted, through mutations of the poly-adenylation signal, mean mRNA expression and transcriptional noise decreased.

I wanted to study this effect on a genome wide scale, measuring the 3’-5’ crosstalk and single cell mRNA expression to measure mean expression and transcriptional noise. If the results from a genome wide analysis are consistent with the results from the in-depth study on model genes, this would suggest that 3’-5’ crosstalk is a common cellular mechanism that increases mean expression and transcriptional noise.

I decided to measure transcriptionally relevant 3’-5’ crosstalk by analysing a publicly available ChIA-PET dataset against RNA Pol2 (Li et al., 2012). This is a method which combines chromatin immuno-precipitation against a protein of interest and chromosome conformation capture. This results in measuring three-dimensional DNA-DNA interactions, which are in proximity of a protein of interest. As I am interested in transcription of mRNA, I used the ChIA-PET dataset with a Pol2 pulldown to find transcriptionally relevant 3’-5’ crosstalk. A similar approach to measure 3’-5’ crosstalk can be found in Grosso et al. (2012).

This technique has a few different versions, for example bridge linker protocol (Tang et al., 2015) or A/B linkers (Fullwood et al., 2009). The dataset I used to measure 3’-5’ crosstalk used the A/B linker protocol, and so this will be explained further (Figure 7). This protocol starts with formaldehyde crosslinking followed by sonication to
fragment the chromatin. Next, the sample is split into two separate aliquots and different linkers (DNA sequences containing the Mme1 recognition site) are ligated onto DNA from each aliquot. The aliquots are mixed together and undergo proximity ligation, followed by Mme1 digestion to release the paired end tags (PETs). The paired end tags are then sequenced to determine the genomic location of each side, which finds areas of the genome that are in close three-dimensional proximity to each other. Chimeric PETs are used to determine non-specific interactions, as these will not come from the same pulldown.

Results

I used ChIA-PET2 (Li et al., 2017) to generate Hi-C style contact matrices at 2000 bp resolution of this publicly available ChIA-PET dataset against RNA Pol2 (Li et al., 2012). I then took the hg19 refgene gene list (O’Leary et al., 2016), and filtered the genes to include only genes with unique gene symbols on chromosomes 1-22 and X, therefore excluding alternatively spliced genes and genes with multiple transcriptional start and end sites (TSS and TES respectively). I also excluded genes shorter than the resolution of these Hi-C contact matrices (2000 bp).
I then calculated the Pol2 mediated interactions between the 3’ and the 5’ ends of a gene by counting the number of reads that correspond to the interaction between bins of the Hi-C matrix that include the TSS and TES. This was inversely correlated with the length of the gene. To counteract this effect, I divided the number of reads for each gene by the average reads from 10,000 genomic intervals of the same length as each gene. These genomic intervals were randomly sampled from across the chromosome. This was transformed by arcsinh $\sqrt{x + 0.5}$, resulting in a variance-stable interaction score (Bartlett, 1947).

I then decided to use a single cell RNA-seq (scRNA-seq) experiment to measure transcriptional noise. I chose to use the results from a droplet sequencing experiment, where the transcriptome of thousands of cells was measured (Klein et al., 2015). These cells were from the same cell line as the ChIA-PET data I interrogated (K562), meaning that the results from scRNA-seq can be compared to the ChIA-PET results.

I downloaded UMI counts data from this scRNA-seq experiment, and the UMI based counts were fitted by Massimo Cavallaro to the simplified two state model (negative binomial), as described in the previous section. Genes were filtered to only include expressed genes (mean UMI count > 0.05) and transcriptional parameters were estimated (e.g. $CV^2$, burst frequency $k_{on}$ and the mean burst size $a/k_{off}$).

We found that log(mean expression) was higher in genes with higher 3’-5’ interaction scores (P<2e-16, linear model, Figure 8). This provides evidence suggesting that 3’-5’ crosstalk of RNA polymerase increases mean expression. This supports the results from the inserted transgenes, and Mapendano et al. (2010).
Figure 8. Plot of mean expression (plotted on a logarithmic axis) vs 3’-5’ interaction score. As 3’-5’ interaction score increases, the average mean expression also significantly increases ($P<2e^{-16}$, linear model). This suggests that crosstalk between the start and the end of a gene increases the gene’s expression.

The burst frequency averaged across all genes determines the noise trend ($CV^2 = 1/\mu + 1/\text{mean}(k_{on})$). I calculated for each gene the vertical distance from this curve on the logarithmic axis (Figure 9 A). The genes above this curve are noisier than predicted, and the ones below this curve are less noisy than predicted. There is significantly higher 3’-5’ interaction score in the genes above the curve ($p<2e^{-16}$, Wilcoxon rank sum). For visualisation purposes, Figure 9 B shows this after removing genes with zero counts corresponding to the interaction between the 5’ and 3’ end.
Figure 9. (A) Plot of $CV^2$ against mean expression, with the noise trend $CV^2 = 1/\text{Mean expression} + 1/\text{mean burst frequency}$ plotted. The genes above the noise trend are marked in blue, while those below are marked in orange. The genes above and below the noise trend are plotted in (B), excluding genes with 0 reads between the 3’-5’ ends. There is significantly higher 3’-5’ interaction score in the genes that have higher noise than expected based on their mean expression ($p < 2e-16$, Wilcoxon rank sum).

The 3’-5’ interaction score (3’-5’ interactions of genes with zero counts were filtered out as they gave little statistical information) was plotted against the transcriptional parameters burst frequency ($k_{on}$) and burst size ($a/k_{off}$). There is a significant negative correlation between 3’-5’ interaction score and burst frequency, and a positive correlation between 3’-5’ interaction score and $\log_{10}(\text{burst size})$ (both $P < 2.2e-16$, linear model, Figure 10 A-B).

Figure 10 (A) Plot of 3’-5’ interaction score against burst size, showing significant positive correlation ($p < 2.2e-16$, linear model) (B) Plot of 3’-5’ interaction score against burst frequency, showing a significant negative correlation ($p < 2.2e-16$, linear model).
These results are consistent between replicates of ChIA-PET and Drop-Seq experiments as well as using different resolutions for the ChIA-PET matrices (Figure 11). I then increased the threshold to only include genes at least double the length of the resolution of the Hi-C contact matrices to avoid possible threshold effects (genes longer than 4000 bp for resolution 2000 bp, genes longer than 10000 bp for resolution 5000 bp and genes longer than 14000 bp for resolution 7000 bp). This also gave consistent results. Furthermore, consistent results were obtained by following the same protocol with a ChIA-PET dataset generated from a different lab (Heidari et al., 2014).

Figure 11. Replicate plots of 3-‘5’ interaction scores between different replicates and different resolutions. Overall, there is very good correlation between different biological replicates and different bin sizes. All genes, without filtering was included for these replicate plots.
Overall, the results presented here provide further evidence that 3’-5’ crosstalk contributes to increased mean expression and increased transcriptional noise on a genome wide scale. This evidence matches the results from inserted transgenes where mutations of polyadenylation signals on two model genes which disrupted 3’-5’ crosstalk were shown to reduce transcriptional noise. This analysis formed part of the publication Cavallaro et al. (2021).
Why do some genes have 3’-5’ crosstalk? Gene Ontology analysis

As genes with higher 3’-5’ crosstalk on average have higher mean expression and transcriptional noise, I wanted to discover what types of genes have high crosstalk, to try and discover a biological reason for why crosstalk exists in the cell. Therefore, I conducted a Gene Ontology (GO) analysis, to see the function of genes with high 3’-5’ crosstalk. I classified a gene as looped if it is in the top 10% of 3’-5’ interaction scores. Across 4 ChIA-PET datasets I analysed, all against Pol2 (3 repeats of K562 and one dataset of MCF7 cells (Heidari et al., 2014; Li et al., 2012)), I consistently see the term “Structural constituent of ribosome” as the most enriched class with adjusted P value being below the adjusted significance threshold of 0.05 for every experiment across the 4 ChIA-PET datasets. No other term was consistently significantly enriched across every dataset.

I analysed the scRNA-seq data presented earlier to discover if these genes that are in this class and are looped behave differently to all other genes. As I expected, they had significantly higher mean mRNA expression than the average across all other genes (P < 2.2e-16, Wilcoxon rank sum), but they were also significantly more variable than all other genes (P = 0.0166, two tailed t-test) (Figure 12 A-B). This is surprising as genes encoding ribosomal proteins are essential for normal cellular processes, are highly expressed in all tissues and have been suggested to be good housekeeping genes (Hsiao et al., 2002). I expected that housekeeping genes would have less variability than other genes, with the rationale that they are essential for normal cellular function, so every cell of the same type will need to transcribe a similar number of them. Indeed, housekeeping genes have been often used as internal controls for expression analysis, with the same rationale. Although it is important to note some housekeeping genes are known to vary across tissues (Eisenberg and Levanon, 2013).
One potential explanation on why these looped genes that are structural components of the ribosome are more variable than other genes could be that these genes need a high expression and increased 3′-5′ interaction score increases mean expression. Perhaps a cell uses 3′-5′ interactions to increase expression, with the trade-off of increased transcriptional noise.

I then decided to compare the looped genes that had the gene ontology term ‘structural components of the ribosome’ to all other looped genes. As I expected, based on their housekeeping function these had a higher mean expression than other looped genes (p<2.2e-16, Wilcoxon rank sum, Figure 13 A). However, the average noise profile is lower for looped genes with the term ‘structural component of the ribosome’ than other looped genes (P=6.84e-4, two tailed T-test, Figure 13 B).
The gene mTOR is known to promote transcription of ribosomal protein genes in response to environmental signals (Mayer and Grummt, 2006). In yeast, this has been found to involve complex mechanisms involving activating other transcription factors (e.g. SPF1, or FHL1) and through controlling chromatin remodelling (Mayer and Grummt, 2006). Activating or turning off this pathway could be used to control transcriptional variation of ribosomal proteins. Similar gene regulatory mechanisms may be used in mammalian cells to control the noise increase derived from increasing looping in these genes, which could explain why these genes have lower transcriptional noise than other looped genes.

Overall, I have discovered that the most enriched function of looped genes (genes in the top 10% of 3’-5’ interactions) correspond to the structural component of the ribosome. These genes are required for life, are expressed in all cells and have been previously referred to as housekeeping (Hsiao et al., 2002). It is therefore surprising that these genes have higher transcriptional noise on average than other genes, due to their basal cellular function. I have speculated that perhaps cells use looping to
increase these genes’ expression, with possibly a disadvantage caused by their increased transcriptional noise. However, these genes do have a lower transcriptional noise profile than other looped genes. This suggests that cells take pathways to combat looping induced transcriptional noise in genes which have the GO term structural components of the ribosome. However, the conclusions from this section may be caveated as they are results from cell lines, which are under selective pressure for fast growth. This is not a selective pressure for humans. Therefore, for cell lines a trade-off between high transcription rates of ribosomal proteins to make new ribosomes required for cell growth (Mayer and Grummt, 2006) and high cell to cell variation may be skewed towards high transcription rates, which may not be the same as in a human.

**Interrogating Pol2 mediated three-dimensional genome structure - Network analysis**

I wanted to investigate the Pol2 mediated three-dimensional structure of the genome, to further interrogate how regions interact with each other. I decided to treat the three-dimensional structure of the genome as a network to achieve this.

I took the Hi-C style contact matrices from two Pol2 ChIA-PET datasets (Li et al., 2012), generated from my analysis pathway from above, and treated this matrix as a network. Here, I defined each node as a 2 kb bin of the genome and links are defined as ChIA-PET interactions between the nodes. Then filtered the matrix to exclude nodes with no interactions and measured how many other nodes each node was linked to. This is known as the degree of the node.

I plotted a figure showing the cumulative fraction of nodes against the degree, both on the logarithmic scale (Figure 14 A-B). I found that this network appeared to exhibit scale-free like behaviour. Here, most nodes interacted with few others but some nodes had many interactions, these high degree nodes are known as hubs. This is shown by the points falling on a straight line on the logarithmic axis, which shows that the degrees follow a power law distribution. Clauset et al. (2009) developed a maximum likelihood-based approach to fit a variety of heavy tailed distributions, and
I used the poweRlaw R package (Gillespie, 2015), which implements this approach to fit the power law distribution to the data. This is compared to a random network model, which was approximated as a Poisson distribution (Barabási, 2016) and also fitted to the data (Gillespie, 2015).

![Figure 14. Plots of cumulative fraction of nodes plotted against degree on the logarithmic axis. Power law and Poisson (approximation of a random network) lines are fitted to this. It can be clearly seen that a power law fits this network better than a random network. (A) Replicate 1. (B) Replicate 2.](image)

These types of networks are robust to failure, as randomly removing a node is unlikely to hit a hub. Interestingly, these network structures are often found in biological systems (Albert, 2005). For example, the yeast protein-protein interaction network appears scale-free (Albert, 2005) and is robust to failure, shown by most gene knock outs having no phenotypic effects (Giaever et al., 2002). Conversely, these networks are susceptible to a targeted attack, where the most highly connected nodes are removed. It has been found that the most highly connected proteins in the yeast proteome are more likely to be essential than less well connected proteins (Jeong et al., 2001).

It appears likely advantageous that the three-dimensional structure of the genome is resistant to a random attack, so that random mutations are unlikely to damage functional three-dimensional interactions. Indeed, three-dimensional interactions of
the genome are remarkably resilient, as short deletions of inter-TAD boundaries does not cause TAD rearrangements (Rodríguez-Carballo et al., 2017). In fact, only a large 400 kb deletion was able to merge neighbouring TADs. The results from my analysis presented here match the results from Sandhu et al. (2012), who also found that the 3D structure of the genome is scale-free.

An interesting feature of these types of networks is how they grow. One method to produce a scale-free network requires new nodes to be added, which are preferentially connected to highly connected nodes. In the Bianconi-Barabasi model, the probability that a link of a new node connecting to an existing one is proportional to the existing one’s degree (Barabási, 2016). This has fascinating implications on how the three-dimensional genome evolved; network growth could perhaps be understood through DNA duplication. However preferential attachment is less clear. Preferential attachment is unlikely to be solely sequenced based, as there must be a limit on binding sites on a defined sequence length. One potential model could be that as a stretch of DNA evolves more binding partners, then the DNA which comes along with these factors provide additional places for further binding to occur.

More recently however, Broido and Clauset, (2019) argue that scale-free networks are rare, and many real-world networks are fitted better by other distributions. I used the poweRlaw R package (Gillespie, 2015) to test if other distributions fit the data better. I fitted four common distributions: discrete power law, Poisson, discrete exponential, and discrete Log-Normal to the nodes. The discrete power law and discrete log-normal were both fitted to the tail of the graph using a Kolmogorov-Smirnov approach (minimising the distance between the data and fitted model CDFs) (Figure 15 A-B).
Visually the Log-Normal distribution fits the data better than the Power-Law distribution, with the other two distributions being poor explanations of the data. Even in the kindest condition to the Power-law (setting the start of tail in a way that minimises the KS distance in the Power-Law, and not the Log-Normal), the Log Normal is still a better fit ($P=1.92e-9$ for replicate 1 and $P=6.67e-9$ for replicate 2, one tailed likelihood ratio test). I can conclude that this chromatin interaction network is not scale-free in the purest sense. However, corrections had previously been made to scale-free network models, which take into account naturally occurring effects in real networks (Barabási, 2016).

One of these corrections, which makes the CDFs deviate predictably from a pure Power-Law is sub-linear preferential attachment (Barabási, 2016). This causes the degree distribution to follow a power law with an exponential cut-off and may reproduce the effects shown here. Molecularly, this could be thought of as steric hinderance on highly connected nodes.

Nevertheless, the distribution of degrees is still heavy tailed with a high number of highly connected nodes (hubs), which has been called ‘far more important than whether it can be fit using a power-law’ (Stumpf and Porter, 2012). I then tested if

---

**Figure 15**. Plots of cumulative fraction of nodes plotted against degree on the logarithmic axis. Power law, Poisson (random network) Exponential and Log-Normal fitted lines are all plotted. Log-Normal fits the data better than Power law for replicates 1 (A) and 2 (B) ($P=1.92e-9$, $P=6.67e-9$ for replicates 1 and 2, one tailed likelihood ratio test).
this heavy tailed distribution of degrees I found is due to insufficient filtering of reads to call an interaction (previously an interaction is called with 1 read), and potentially spurious reads may affect my results. I chose two different thresholds, 5 and 50 reads to call an interaction and in both cases the distribution is still heavy tailed in both replicates.

I wondered if these highly connected nodes contained functional genetic elements. I therefore tested if the hub nodes contained more genes and enhancers than the less highly connected nodes.

I used published lists of enhancer locations from Enhancer Atlas 2.0 (Gao and Qian, 2020) specific to K562 cells and gene locations (O’Leary et al., 2016) to find where functional genetic elements overlapped with nodes. I then defined a hub as nodes with the highest degree (top 1%) which corresponded to a degree of 15. I then determined if hubs and non-hubs overlapped with at least one enhancer and with at least one gene (Table 7 A-D).

| Table 7. Tables showing if nodes in hubs (top 1%) overlap with genetic features more than nodes in non-hubs (lower 99%). (A) Replicate 1 looking at enhancers. (B) Replicate 2 looking at enhancers. (C) Replicate 1 looking at genes. (D) Replicate 2 looking at genes. In all cases, hubs overlapped more with genes and enhancers than non-hubs (P<2.2e-16 for all, Chi-Squared) |
On average, the hubs overlapped with enhancers ($p<2.2e^{-16}$, chi-squared) and genes ($p<2.2e^{-16}$, chi-squared) significantly more than the non-hubs, in both replicates. This result was maintained for enhancers ($P<2.2e^{-16}$, chi-squared) as well as genes ($P=1.01e^{-13}$, chi-squared) when using a higher threshold of 5 reads required to call an interaction. I can therefore conclude that enhancers and genes are more often found in hubs. I found this interesting, as it can provide an additional evolutionary explanation of this scale-free like network structure (along with being robust to failure). For example, the presence of enhancers in these hubs could mean that a cell could efficiently use the same enhancer to regulate many processes, wasting less effort on producing additional enhancers. This is supported by the fact that enhancers can simultaneously activate multiple genes, e.g. Fukaya et al. (2016). Furthermore, the presence of genes in these hubs points to their ability to be regulated by multiple distal DNA loci. It would be interesting to interrogate hubs using ChIA-Drop (Zheng et al., 2019), to test if hubs are areas where multiple elements come together simultaneously, or if these interactions are temporally separated.

Here, I have shown that the Pol2 mediated three-dimensional interaction network appears to be scale-free, with a correction such as non-linear preferential attachment. These networks are typified by being robust to random attack, but susceptible to targeted attacks, and are often found in biological systems. I have also found that hubs in these networks are more likely to overlap with both genes and enhancers, which could provide an additional evolutionary explanation for using this structure, potentially increasing efficiency of regulation.

**How does polymerase occupancy contribute to transcriptional noise?**

I wanted to measure 3′-5′ crosstalk in a different manner to ChIA-PET as this would give me more confidence that the genome wide results presented above are not a technical artifact of the ChIA-PET protocol or my analysis. I reasoned that the Pol2 occupancy at the 5′ and 3′ end of a gene could be used as a proxy for 3′-5′ crosstalk. This is because if a protein connects two DNA loci according to ChIA-PET, the protein is likely to show occupancy at the same loci per ChIP-Seq.
Polymerase occupancy can be measured through ChIP-Seq. This technique involves fixing DNA-protein interactions, sonication, and an antibody pulldown of the protein of interest. This is followed by library preparation and sequencing of DNA associated with this protein. ChIP-Exo is a complementary technique, which follows a similar protocol but contains an exonuclease digestion stage to improve the resolution of ChIP-Seq.

This means ChIP-Seq data could potentially be used as another way to measure 5’ to 3’ crosstalk by a particular protein factor, in this case Pol2. I therefore analysed a ChIP-Seq and ChIP-Exo experiment (McHaourab et al., 2018) to see if 3’-5’ crosstalk measured in this way also contributed to transcriptional noise. This experiment used an antibody against the N-terminal domain of RPB1, and so will pull down total Pol2.

Indeed, 3’-5’ crosstalk has been measured in a conceptually similar fashion before (Grosso et al., 2012). Here the authors chose to measure crosstalk from measuring phosphorylated Ser-5 on the Pol2 CTD at the 3’ end of a gene. Ser-5 phosphorylation is known to accumulate at the start of transcription and decrease through elongation (Mayer et al., 2010) so the authors postulate that 3’ accumulation of phosphorylated Ser-5 Pol2 is likely to be caused by 3’-5’ crosstalk. They found that genes with crosstalk are shorter and have higher mean expression. They also found that these genes with 3’ Ser-5 phosphorylation are more likely to contain the consensus poly-adenylation signal A2TA3, and genes without 3’ Ser-5 phosphorylation are more likely to contain weaker poly-adenylation signals. This supports our results on studying the inserted model genes HBB and HIV env where mutation of the poly-adenylation signal caused less 3’-5’ crosstalk, reduced mean expression and reduced transcriptional noise. However Ser-5 phosphorylation decreases gradually through a gene which results in shorter genes having more phosphorylated Ser-5 at the 3’ end (Hsin and Manley, 2012). This could bias results towards shorter genes based on using Ser-5 phosphorylation at the 3’ end to show 3’-5’ crosstalk.
I attempted to discover if crosstalk between the 5’ and the 3’ end of a gene, as measured by ChIP-Seq contributes to transcriptional noise. I downloaded RPKM normalised coverage files of total Pol2 ChIP-Seq and ChIP-Exo experiments (McHaourab et al., 2018). I calculated the average read coverage at the 5’ end (TSS +/- 250 bp) and the 3’ end (TES +/- 250 bp) at all genes using the program bwaverageoverbed (Kuhn et al., 2013).

I then used the same estimates for transcriptional noise from earlier in this thesis, and from Cavallaro et al. (2021). Briefly, I calculated the vertical distance \( (v) \) from the predicted noise trend (Figure 16). This is because the cell lines in the ChIP-Seq and ChIP-Exo experiment (K562) are the same as those we used previously to calculate transcriptional noise.

I filtered the genes by removing those shorter than 500 bp, as otherwise the measurements of Pol2 occupancy at the 5’ and 3’ end would overlap. I also removed genes which had multiple TSS or TES, and those which had 0 reads at the 5’ or 3’ end of the gene. I also subsampled each dataset to select protein coding genes.
I constructed 2 linear models where mean expression as well as distance from the predicted curve (transcriptional noise) is predicted by the average coverage at both ends of the gene (Table 8).

<table>
<thead>
<tr>
<th>Sample: Log(mean expression)</th>
<th>Log(TSS coverage) predictor P value</th>
<th>Log(TES coverage) predictor P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChIP-Seq protein coding</td>
<td>0.444</td>
<td>0.739</td>
</tr>
<tr>
<td>ChIP-Seq all genes</td>
<td>0.000224</td>
<td>0.029051</td>
</tr>
<tr>
<td>ChIP-Exo protein coding</td>
<td>0.0398</td>
<td>0.7326</td>
</tr>
<tr>
<td>ChIP-Exo all genes</td>
<td>3.67e-07</td>
<td>0.382</td>
</tr>
</tbody>
</table>

Table 8. (A) Table showing the predictor P values for a linear model predicting log(mean expression) by log(coverage) at the Transcriptional Start site (TSS) or Transcriptional End Site (TES), for a ChIP-Exo or ChIP-Seq experiment – looking at either all genes or all protein coding genes. Overall, it appears that expression is predicted by coverage at the TSS and not the TES. (B) Table showing the predictor P values for a linear model predicting transcriptional noise (v, distance from the noise trend) by log(coverage) at the Transcriptional Start Site (TSS) or Transcriptional End Site (TES), for a ChIP-Exo or ChIP-Seq experiment – looking at either all genes or all protein coding genes. Overall, it appears that Transcriptional Noise is predicted by coverage at the TSS not the TES.
At first glance, whilst not every P value is below the threshold of <0.05, it appears that overall, Pol2 occupancy at the 5’ end of a gene contributes to mean expression and transcriptional noise. When 5’ occupancy increased, mean expression and transcriptional noise also increased. However, Pol2 occupancy at the 3’ end did not appear to contribute towards either.

Pol2 often accumulates at the 5’ end of genes, which is believed to be caused by recruited Pol2 pausing before entering productive elongation (Adelman and Lis, 2012), and so high 5’ occupancy can be an indicator of high pausing. This interpretation of my results matches the results from Zhang et al. (2021) who found that higher pausing times and pausing frequencies contributed towards transcriptional noise. Although, abortive transcription could also explain high Pol2 occupancy at the 5’ end, see Part 2, Chapter 4 and Zhang et al. (2021).

However, when adding an interaction term into the model to measure crosstalk between 5’ and 3’ of a gene, an interesting pattern emerges (Table 9). Overall, I found that the interaction between coverage at both ends of a gene was below the significance threshold of 0.05 in most samples for both mean expression and transcriptional noise, which means both the mean expression and transcriptional noise are predicted by the product of coverage at both ends of a gene. In every case this direction was positive, meaning that transcriptional noise and mean expression is increased when there is a high Pol2 occupancy at both the 5’ and 3’ end of a gene.

<table>
<thead>
<tr>
<th>Sample: Log(mean expression)</th>
<th>Log(TSS coverage) predictor P value</th>
<th>Log(TES coverage) predictor P value</th>
<th>Interaction term predictor P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChIP-Seq protein coding</td>
<td>0.4725</td>
<td>0.1578</td>
<td>0.0461</td>
</tr>
<tr>
<td>ChIP-Seq all genes</td>
<td>0.8451</td>
<td>0.0981</td>
<td>2.48e-05</td>
</tr>
<tr>
<td>ChIP-Exo protein coding</td>
<td>0.221</td>
<td>0.233</td>
<td>0.201</td>
</tr>
<tr>
<td>ChIP-Exo all genes</td>
<td>0.00187</td>
<td>0.12687</td>
<td>0.00106</td>
</tr>
</tbody>
</table>

### A

<table>
<thead>
<tr>
<th>Sample: Transcriptional</th>
<th>Log(TSS coverage) predictor P value</th>
<th>Log(TES coverage) predictor P value</th>
<th>Interaction term predictor P value</th>
</tr>
</thead>
</table>

### B

73
Table 9. (A) Table showing the predictor P values for a linear model predicting log(mean expression) by log(coverage) at the Transcriptional Start site (TSS) or Transcriptional End Site (TES), as well as including an interaction term. This is for a ChIP-Exo or ChIP-Seq experiment – looking at either all genes or all protein coding genes. Overall, it appears that expression is predicted by the interaction of coverage at the TSS and the TES. (B) Table showing the predictor P values for a linear model predicting transcriptional noise (v, distance from the noise trend) by log(coverage) at the Transcriptional Start site (TSS) or Transcriptional End Site (TES) as well as including an interaction term. This is for a ChIP-Exo or ChIP-Seq experiment – looking at either all genes or all protein coding genes. Overall, it appears that Transcriptional Noise is predicted by the interaction between coverage at the TSS and the TES.

One possible confounder for these results could be that coverage at the 5’ end and 3’ end of a gene correlate with each other (Pearson’s R² = 0.23 for ChIP-Exo all genes dataset, on the logarithmic scale, Figure 17) and this multicollinearity is known to be a problem with linear regression (J Neter, MH Kutner, CJ Nachtsheim, 1996). One formal way of investigating this is to use the Variance Inflation Factor (VIF) to quantify the severity of multicollinearity. One suggested threshold is if VIF is greater than 10, then multicollinearity is classified as severe (J Neter, MH Kutner, CJ Nachtsheim, 1996). In all linear models, VIF for every predictor variable is less than this threshold, meaning that multicollinearity is not severe, and linear regression is appropriate.
This provides further evidence towards the hypothesis that crosstalk between the 5’ and 3’ end of a gene contributes towards transcriptional noise and mean expression as there is a significant interaction between the coverage at both ends of a gene.

Summary

Here, I measured transcriptionally relevant 3’-5’ crosstalk on a genome wide scale by analysing ChIA-PET datasets with a Pol2 pulldown (Heidari et al., 2014; Li et al., 2012). I combined this with calculations of transcriptional noise from a scRNA-Seq dataset (Klein et al., 2015). I found that genes with higher crosstalk had on average higher transcriptional noise, mean expression and burst size. These genes also had lower burst frequency. This is further supported by analysing a ChIP-Seq, and ChIP-Exo experiment, also against Pol2 (McHaourab et al., 2018) where I found that there is a significant interaction term in the linear model between coverage at the 5’ and 3’ end of a gene and transcriptional noise. Taken together with the results from Chapter

Figure 17. Coverage of the Transcriptional Start Site (TSS) plotted against the Transcriptional End Site (TES), where each point represents a gene from the ChIP-Exo dataset. There is correlation between coverage at the TSS and the TES (Pearson’s $R^2 = 0.23$)
1, I can conclude that 3’-5’ crosstalk contributes towards transcriptional noise, which matches the hypothesis I am testing.

Further analysis of genes with very high 3’-5’ crosstalk (top 10% of 3’-5’ interaction score) found that the most enriched class was ‘structural component of the ribosome’. These genes on average had higher transcriptional noise and mean expression than all other genes. However, they had lower transcriptional noise than other genes with very high 3’-5’ crosstalk. I speculated that transcriptional noise could be disadvantageous for housekeeping genes, and perhaps cells use other mechanisms to reduce transcriptional noise in these genes, compared to other genes with very high 3’-5’ crosstalk.

Upon treating the three-dimensional genome as a network, I found that this networks’ degree distribution appeared scale-free (with a non-linear preferential attachment correction). Here, most nodes had very few interactions, but some had many more than predicted by a random network. I found that the nodes with many interactions overlapped significantly with genes and enhancers – potentially providing an evolutionary explanation for this network structure.

In the next part of this thesis, I develop a method to label and track Pol2 from a specific genomic locus. I then use this method to measure the local recycling of Pol2 on a model gene (HBB) where crosstalk can be disrupted by a poly-adenylation point mutation. This is used to test my hypothesis that the mechanism of 3’-5’ crosstalk contributing to transcriptional noise is through polymerase recycling causing a burst of transcription.
Part 2: A novel method to label and track Pol2 from a specific genomic locus

Chapter 1: Developing the method

Introduction

Previous research looking at mutations in polyadenylation sites at the 3’ end of a gene has suggested that RNA polymerase can be recycled (Mapendano et al., 2010). Here the authors took cell lines which expressed the tetracycline repressor protein as well as containing single copies of Human Beta-Globin (HBB) or HIV env genes, in HEK293 cell lines. These genes (HBB and HIV env) contain two copies of the Tet operator after their CMV promoter and expression is induced by the addition of tetracycline.

These integrated genes either have WT polyadenylation signal (WT), or a point mutation A\textsubscript{2}TA\textsubscript{3} -> A\textsubscript{2}GA\textsubscript{3} (pA-). Crucially, after tetracycline induction, this point mutation at the 3’ end of a gene resulted in reduced Pol2 at the promoter, which is suggested to be due to the inhibition of recycling (Figure 18) (Lykke-Andersen et al., 2011; Mapendano et al., 2010).

![Figure 18](image.png)

*Figure 18. A plot comparing Pol2 occupancy at the HBB promoter compared to the GAPDH promoter in the WT line and the pA- line (point mutation at poly adenylation signal), after different timings of HBB induction (250ng/ml tetracycline). The mutant (dark grey) loses polymerase accumulation on the promoter, whereas it is maintained on the wild type (light grey). Figure from (Mapendano et al., 2010) and reproduced under license.*
Published work from our laboratory (Part 1, Chapters 1-2) has shown that this crosstalk between the 5’ end and the 3’ end of genes contributes to transcriptional noise (Cavallaro et al., 2021). We found that the genes with WT poly-adenylation signals had higher expression levels and increased transcriptional noise, suggesting that polymerase recycling contributes to both.

A microscopic model of transcription on these two cell lines, implemented by Massimo Cavallaro, further supports this (Figure 19, Cavallaro et al., 2021). We found that reducing the recycling rate of this model (l, Figure 19) could reproduce the effects of the poly-adenylation signal point mutation. Whilst this is strong evidence that polymerase recycling contributes to transcriptional noise, it is also indirect. Furthermore, increasing the rate of recruitment to the pool of polymerase (γ) when the gene is in the active state (Green, Figure 19) was not investigated. Mechanistically this could involve the Tet repressor reducing the efficiency of GTFs to bind DNA. When the gene is activated, the Tet repressor leaves and could increase the efficiency of GTF binding. This increase could be different between the WT and pA- line and may also reproduce the effects seen in the experimental system. I therefore wanted to develop a system that more directly measures polymerase recycling.

Figure 19. Figure from Cavallaro et al. (2021), Image made by Massimo Cavallaro. A microscopic model of transcription of the system. It may be possible that the rate γ could be increased when the gene is active (green compartment), and this increase could be different between the WT and the pA- line, which could reproduce the effects seen in the experimental system.
To test the hypothesis that polymerases are recycled in these cell lines I have developed a system, originally envisioned by Andrew Bowman, where Pol2 is labelled through biotinylation at the HBB locus. Labelled Pol2 is then tracked by purifying against the biotinylation tag, and measuring the DNA associated with it.

This system works by transfecting a cell line which contains a gene with a Tet operon right after the transcriptional start site with two fusion proteins that I designed; the Tet repressor fused to a biotin ligase (TetR-BirA) and the Avitag (a biotin acceptor peptide) fused to the main Pol2 subunit RPB1 (Avi-RPB1). Both fusions were synthesised and cloned into an expression plasmid by an external supplier. This system works in a conceptually similar fashion to BioID (Kim et al., 2016; Roux et al., 2012), but uses a non-promiscuous biotin ligase that is designed to specifically only biotinylate Avi-RPB1. Schlissel and Rine (2019) also take a similar approach to biotinylate and track histones.

In these cell lines containing the Tet operon and without the presence of tetracycline, the TetR binds to the Tet operon and blocks transcription of the downstream gene. In my system, the TetR-BirA fusion protein works in a similar fashion, but also biotinylates Avi-RPB1 at the Tet operon due to their close physical proximity. Avi-RPB1 is then released through the addition of tetracycline, which causes TetR-BirA to unbind and allows Pol2 to proceed with transcription. This Avi-RPB1 can then be tracked through pulldown of the biotin tag to find out where Pol2 travels after it is released from a specific genetic location (see Figure 20 for a schematic). I used a HEK293 cell line, with a single genomic integration of HBB, under the control of the Tet operon for this work. This is one of the cell lines used in Cavallaro et al. (2021) and Mapendano et al. (2010).

I hoped to be able to check for recycling by seeing a Biotin-Pol2 signal which would be maintained on HBB after tetracycline induction. I planned to see if Biotin-Pol2 is maintained on HBB more on the WT line than pA- line, which would directly demonstrate if recycling is altered on these lines (Figure 20).
Figure 20 A schematic of the experimental protocol, the TetR-BirA fusion protein biotinylates Avitag-RPB1 in close proximity to the Tet operon. ChIP-Seq with a biotin pull down after tetracycline induction is then used to determine the location of Pol2 that was previously close to the Tet operon. A time course ChIP experiment could then resolve if the same polymerase stays on the gene, which could be a direct measurement of recycling.

Results

My first test of this technique was to transfect approximately 700,000 HEK293 HBB WT cells in a 6-well dish with either a plasmid encoding AviTag-RPB1, a plasmid encoding TetR-BirA or transfecting with both plasmids (1000 ng, 500ng, 250ng and 100ng total DNA content, split evenly between the plasmids) as well as supplementing the media with 50 µM biotin, as recommended by Roux et al. (2012). 1000ng total DNA transfected resulted in cell death, so this sample was removed from the experiments. I then proceeded with a western blot using an anti-biotin HRP antibody with a biotinylated protein ladder as a positive control.
I expected to see no biotinylated Avi-RPB1 in any of the lanes which only had one plasmid transfected and biotinylated Avi-RPB1 in the 3 lanes which had both plasmids transfected. I saw biotinylation in the 500ng and 250ng lane, but not the 100ng, at the right size for RPB1, suggesting that TetR-BirA does biotinylate Avi-RPB1 (Figure 21). As I saw maximum biotinylation when 500ng of total DNA was transfected, I transfected 500ng in future experiments of the same cell numbers.

Figure 21. A western blot of different concentrations of DNA transfected. There is only 1 band at the correct size when both plasmids are transfected. This suggests that the system is specific as Pol2 is not biotinylated when only one plasmid is transfected. The largest band was seen when 500ng of plasmid were transfected, and this concentration was used for future experiments.

My next experiment was to check that this is due to true proximity based biotinylation between the TetR-BirA on the Tet operon and Avi-RPB1 or due to freely diffusing TetR-RPB1 biotinylating freely diffusing Avi-RPB1. I did this by repeating the experiment on a cell line that did not have the Tet operon. While HEK293 cells without the Tet operon would have been a more consistent model, our laboratory did not own this line and so I used HeLa cells. If biotinylation could only occur on the Tet operon, then these cells will not have any biotinylation, demonstrating the specificity of my system.
This time I used streptavidin conjugated to HRP. This was because I was planning to use streptavidin coated beads for my ChIP pulldown, and I wanted to check if streptavidin–HRP also detected biotinylated Avi-RPB1. I used four samples, all with 500 ng DNA transfected, (250 ng TetR-BirA, 250 ng Avi-RPB1), two with biotin enriched media (50 μM) and two without biotin enriched media. I only saw bands of the correct size corresponding to biotinylated Pol2 in the cells containing the Tet operon, and not in the cells without the Tet operon (Figure 22 A). I also saw higher background, with a smear in the lanes containing cells grown in biotin enriched media. I suspected this was due to the presence of residual biotin in these lanes, and as biotin enrichment was not required to biotinylate Avi-RPB1, it was not added for future experiments.

I also repeated the western blot to see if both plasmids are required to be transfected when using streptavidin HRP to stain the blot (Figure 22 B). This was the case, matching the results from earlier.

Overall, from these experiments I have found that Avi-RPB1 gets biotinylated only when both TetR-BirA is transfected and when the cell line contains the Tet operon. I have also found that adding biotin enrichment to the media is not required and may increase contamination, demonstrated by the smear on the western blot. This shows
that biotinylation occurs as per expectations with specific biotinylation of Avi-RPB1 at the Tet operon by the fusion protein consisting of TetR and BirA.

Because the largest subunit of Pol2 has been modified, both through the addition of the AviTag, and through biotinylation of this tag, I wanted to check that this fusion protein behaves like a normal Pol2. Similarly, as the Tet repressor is fused to a biotin ligase, I also wanted to check if it still behaves similarly to a normal Tet repressor and holds Pol2 at the Tet operon, until tetracycline is added, upon which it dissociates from the Tet operon to allow polymerase progression.

I decided to test these two questions simultaneously by conducting ChIP-Seq experiments with an anti-biotin pull down, before and after tetracycline induction. I had two hypotheses; 1) If the anti-biotin ChIP-Seq appears similar to a normal Pol2 ChIP-Seq then this indicates that the fusion protein behaves like normal Pol2. 2) If the reads at other genes (not containing Tet operon) are higher after tetracycline induction than before tetracycline induction, then this suggests biotinylated Pol2 is released from the Tet operon following tetracycline induction.

I therefore proceeded to optimise my chromatin pulldowns for the ChIP-Seq experiments. Ideally the size distribution of DNA fragments should be between 100 and 500 bp for ChIP-Seq experiments (Sullivan and Santos, 2020), so this is the general range I was aiming for. I optimised this by sonicating formaldehyde fixed cells for multiple cycles of sonication (5 mins of 30s on 30s off at high power). I found the optimum size distribution was at 6 x 5 min cycles (Figure 23).
ChIP-Seq experiments typically require 1-10 million cells for a good signal to noise ratio (Kidder et al., 2011), and the previous experiments were conducted inside a 6 well dish which only reaches approximately 1.2 million cells at confluency. I decided to scale up the experiment, working with a 100mm dish and initially seeding approximately 2.5 million cells which would then be transfected the next day. Using a rough estimate of a doubling time of 24 hours, I expected this to result in around 10 million cells after 2 days (cells left overnight to seed, and then left overnight after transfection), which should be sufficient for a good signal to noise ratio. As I increased the surface area of the dish used, I also increased the mass of plasmid transfected correspondingly to 2.96 µg total DNA content, 1.48 µg of each.

I then conducted a trial run of library preparation using chromatin pulled down with streptavidin linked beads, which appeared to work well (Figure 24).

Figure 23. An agarose gel showing the effects of different sonication times on DNA length distribution. Overall, most DNA was between 500 bp and 100 bp with 6x 5 min cycles of 30s on 30s off high power.
I prepared samples for a time course ChIP-Seq experiment to attempt to answer the questions defined above. I prepared samples from the HBB WT and pA- cell lines, described above, for the following time points after tetracycline induction. 0 (no induction), 30 seconds induction, 60 seconds induction, 5 min induction and 1 hour induction. No input control was performed, as I wanted to compare the samples to 0 (no induction).

After library preparation, the samples were sequenced using NovaSeq 150 bases, paired end. I constructed two custom reference genomes by inserting the HBB gene loci into its correct location (Cavallaro et al., 2021) for both the WT and the pA- gene
(chromosome 12, position 1,223,244). I then adjusted the annotation files accordingly.

Simply, the bioinformatic pipeline involved quality checking with fastqc (Andrews, 2010), quality and adaptor trimming with Trim Galore (Krueger, 2021) and alignment with bowtie2 (Langmead and Salzberg, 2013). Metagene plots were constructed (Figure 25) with counts per million normalisation (CPM) ignoring reads from plasmids, sex chromosomes and the mitochondrial genome using the deeptools package (Ramírez et al., 2016).

**Figure 25. Counts per million normalised metagenes after differing times of HBB induction.** (A) WT samples (B) pA- samples. Both of these look like a typical Pol2 metagene, with a peak at the transcriptional start site. However, surprisingly the baseline CPM seems to vary significantly between different times of tetracycline induction.
Both samples at time point 0 (with no tetracycline induction) showed signs of polymerase release as the metagene profiles have a peak at the TSS. Surprisingly, this peak is larger than some of the samples after tetracycline induction. Furthermore, the baseline values outside genes appeared to vary between repeats, which was again surprising. A final issue with this dataset is that there appears to be a spike towards the transcriptional end site. This spike is likely to be artefactual as it is much thinner than would be expected from a true peak. There also appears to be a small peak approximately 1.5 kb before the peak at the TSS, which is not observed in later experiments, or published Pol2 ChIP-Seq experiments (Figure 28, (Wu et al., 2020)). I have not been able to determine the cause of this peak.

I also noticed a high proportion of DNA pulled down seemed to align to the mitochondrial genome. The literature indicates that the mitochondria contain endogenously biotinylated proteins (Hollinshead, Sanderson and Vaux, 1997). DNA associated with these could also be pulled down, as well as biotinylated Avi-RPB1. I thought that this might be useful as an internal control to normalise the ChIP-Seq reads against as I would not expect these to change between conditions, potentially reducing the difference in baseline between timepoints. I normalised my samples against the aligned reads in the mitochondrial genome and constructed metagenes profiles again (Figure 26). This did not seem to reduce the difference in baseline.
As these samples have both release from the Tet operon before induction as well as differences in baseline coverage between samples, I decided to improve my protocol.

I thought that the release from the Tet operon could be due to trace presence of tetracycline present in the media. A likely source of tetracycline in the media could be foetal bovine serum (FBS) which is added to the cell culture media. This may be contaminated with tetracycline if the mother was being treated with antibiotics.

Figure 26. Mitochondrial scaled metagenes (10000/total mitochondrial reads) after differing times of HBB induction. (A) WT samples (B) pA- samples. Both of these look like a typical Pol2 metagene, with a peak at the transcriptional start site. Scaling by mitochondrial DNA content does not seem to reduce the difference in baseline between different samples.
Another possibility could be that too high concentrations of plasmids being transfected caused off-target effects.

The difference in baseline coverage could be due to differences in transfection efficiency. For example, if the transfection efficiency is poor in one sample, then less Pol2 will be biotinylated. This could result in decreased signal to noise ratio, decreasing the peak size at the TSS and thereby increasing the baseline.

I discovered that the artefactual peak shown in figure 25 and 26 was due to three regions which had an abnormally high number of reads (Table 10). I therefore adjusted the custom reference genome to blacklist these regions (replace the base pairs with N). This will result in no DNA aligning to this region and will remove the artefactual peak in the metagenes.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Start region</th>
<th>End region</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>1225144</td>
<td>1225295</td>
</tr>
<tr>
<td>12</td>
<td>1225390</td>
<td>1225539</td>
</tr>
<tr>
<td>12</td>
<td>1229973</td>
<td>1230123</td>
</tr>
</tbody>
</table>

*Table 10. A table showing the regions of the genome which had many artefactual reads. These regions will be blacklisted.*

I also blacklisted the endogenous HBB gene, as this is not expressed on the RNA level in HEK293 cells (Uhlén et al., 2015, Figure 27). It is therefore likely that most ChIP-Seq reads which align to HBB are likely to come from the inserted transgene, as Pol2 is unlikely to be present on a non-transcribing gene. Blacklisting the endogenous HBB gene will mean that reads aligning to HBB will not be called as multi-mapping reads and discarded, as they will only align to the insert on chromosome 12.
I therefore conducted 3 independent anti-biotin ChIP-Seq experiments using verified Tet-free FBS. I transfected cells with plasmids encoding the fusion proteins at 3 different concentrations of each plasmid (1.48µg, 148ng and 14.8ng), to test the idea that too high concentrations of plasmids may cause off target effects. I used an empty px458 plasmid to keep the total DNA mass transfected constant. This time I used a longer time period of tetracycline induction (24 hours) as I wanted to ensure HBB was maximally induced.

I also adjusted my protocol for transfection to reduce variability in transfection. Simply, I transfected the cells all together and the next day split these into two dishes. As the cells from each dish were transfected together, there should be no variation in transfection efficiency between the induced and the not induced cells. As the cells were left for longer after induction (an extra 24 hours) I correspondingly reduced the initial seeding density to 1 million cells. After sequencing, I constructed metagenes (CPM) using the same protocol as before (Figure 28 A-C).

Figure 27. A chart of normalized transcripts per million of the HBB gene for multiple different cell lines, sorted by expression. This chart is from the Human proteome atlas (Uhlén et al., 2015). The HEK 293 cell line which the cell lines used in our experiment derive from is marked. As there is no expression of endogenous HBB, it is likely that reads aligning to HBB should be mapped to the ectopic gene.
Again, there appears to be polymerase release from the Tet operon before tetracycline induction, suggesting that using Tet free FBS and reducing plasmid concentrations does not completely stop polymerase release. However, it does appear that tetracycline does increase the release – as demonstrated by the increased peak at the TSS after 24 hours induction on the metagene plots. This is illustrated through constructing metagene plots comparing the ratio of 24-hour induction to without induction (Figure 28 D). This demonstrates that the fusion protein TetR-BirA works as expected, and adding tetracycline allows biotinylated Pol2 to spread out across the genome. Interestingly, the low and medium transfection
metagenes have a small drop towards the TES (Figure 28 A-B), resembling metagenes found in the literature (e.g. Ramírez et al., 2014), whereas this is not present in the high transfection metagene (Figure 28 C), however this may be seen when looking at the ratio (Figure 28 D).

Furthermore, the issue with the baseline being different between different time points between samples appears to be solved, probably through reducing the variability in transfection efficiency.

Finally, the artefactual spike towards the TES has now been removed, due to blacklisting the offending regions on chromosome 12. Overall, this shows that the technical problems with the first experiments have been fixed, and this experiment can be analysed in more detail.

I analysed a Pol2 ChIP-Seq experiment using a similar cell line (HEK293, without HBB insertion (Wu et al., 2020) Accession number: SRR10349586) using the same analysis pipeline (except for single end alignment). I constructed a metagene plot as before and found that the anti-biotin pulldowns resemble Pol2 ChIP-Seq (Figure 28 E).

I also inspected regions on IGV and compared these regions to Pol2 ChIP-seq data. I found that while there appeared to be more noise in my pulldown, regions with high Pol2 ChIP-seq reads appeared to be the same with my data (Figure 29 A).

I also inspected the CPM normalized coverage at the inserted gene, HBB for all my pulldown samples (Figure 29 B). I expected to see a large peak of biotinylated Pol2 at the HBB promoter before tetracycline induction, which would dissipate after induction. Surprisingly, I found minimal coverage in HBB for all samples, both before and after transfection. There did seem to be more reads towards the end of the HBB gene in the datasets with 1.48 µg of each plasmid transfected (Figure 29 B). However, this peak appeared artefactual as closer inspection revealed that almost every read aligning to this region did not have a successfully mapped mate. Zooming in on the Tet operon (Figure 29 C) shows very low coverage at this location in all samples,
which did not agree with my expectation of a large peak of biotinylated Pol2 before tetracycline induction. There is a very small number of reads here before polymerase induction for the 1.48 µg and the 148 ng samples, which may suggest the presence of Biotin-Pol2 held at the Tet operon, which dissipates upon induction, however the very small number of reads here makes it difficult to have firm conclusions. This may be caused by a high degree of abortive transcription at HBB (Part 2, Chapter 4).
Overall, even considering the artefactual peak in HBB, the sample with 1.48µg of each plasmid transfected (high) appears to perform the best. This is because it has the largest metagene peak after tetracycline induction, which suggests more Pol2 is biotinylated with this sample. Therefore, the following analysis proceeded with this sample.

I used MACS2 (Zhang et al., 2008) to call ChIP-Seq peaks of the Pol2 dataset along with my sequencing dataset and I used R packages ChIPpeakAnno and genomation (Zhu et al., 2010; Akalin et al., 2015) to analyse these peaks. The threshold to determine a peak was left at the MACS2 defaults. In both cases the peaks from my sequencing dataset significantly overlapped with the Pol2 ChIP-Sequence dataset (P = 9.2e-21 for 0hours, P = 7.9e-33 for 24 hours, Hyper geometric test) as well as each other (P = 2.6e-300, Hyper geometric test) (Figure 30 A). The universe for this Hyper geometric test was estimated using ChIPpeakAnno’s default method. Here the Human genome size multiplied by 0.03 (approximating 2% coding, 1% regulatory regions) and divided by the median peak width, which resulted in approximately 55,000 possible binding regions for Pol2. It should be noted however that my sequencing peaks appear to be a subset of the Pol2 ChIP-Seq peak, this can perhaps be ascribed to lower sequencing depth of my samples (~7 million reads vs 30 million reads) or that few polymerases are labelled at the Tet operon. Manual inspection of the samples using IGV also suggests the presence of uncalled peaks on TSS of many genes, possibly due to the lack of sequencing depth.
I then determined the location of peaks with respect to genes. I selected the nearest gene to each peak, within a distance of 5000 bp. I then established if the peak was upstream, overlapped the start of a gene, was inside a gene, covered the whole gene, overlapped, or was downstream of a gene (Figure 30 C). I found that 24 hours after tetracycline induction appear to be more like a typical Pol2 ChIP-Seq experiment, with 13.2% of peaks overlapping the start of genes compared to only 1.83% overlapping the start of genes without tetracycline induction. This is compared to 43.2% of peaks which overlap the start of genes with Pol2 ChIP-Seq. I filtered these peaks to only look at uniquely Tet induced peaks, i.e., peaks detected in the Tet induced samples, but not in the no induction samples. I found that 64% of uniquely Tet induced peaks overlapped with a gene (peaks overlapping start, end, inside or the whole gene), which was even more similar to Pol2 ChIP-Seq (62%). This may suggest that some non-genic regions of the genome are stably bound by other biotinylated proteins, which may be detected by looking at peaks without tetracycline induction. For example, some histones have been shown to be naturally biotinylated (Kuroishi et al., 2011), though this is a very rare event, thought to be less than 0.001% of human histones (Kuroishi et al., 2011). This may also be supported by the fact that the peaks from without induction and with tetracycline induction overlap extremely well (P = 2.5e-300, Hyper-geometric test).
Figure 30. Analysis of MACS2 called peaks for normal Pol2, and Biotin-Pol2 ChIP-seq before and after induction of the 1.48 µg DNA transfected sample (high) (A) Venn diagrams showing the overlap of MACS2 called peaks. There is significant overlap between all called peaks (hyper geometric test) (B) Boxplot showing normalized transcripts per million for genes associated with MACS2 called peaks. (C) Pie charts showing where, in comparison to genes, these peaks are located.
I also determined if the genes associated with these peaks had differential expression. I downloaded normalized transcripts per million values (nTPM) from the human proteome atlas (Uhlén et al., 2015) for HEK293 cells and filtered to only include expressed genes (nTPM >= 1). I found no significant difference between expression of genes associated with Pol2 peaks and biotin Pol2 peaks without tetracycline expression (P = 0.76, two-tailed Wilcoxon Figure 30 B), but I found the expression of genes associated with biotin Pol2 peaks was significantly higher than genes associated with Pol2 peaks (P = 0.024, two-tailed Wilcoxon). This suggests that only a small proportion of Pol2 is biotinylated, as highly expressed genes are likely to have a lot of (biotinylated) Pol2, meaning MACS2 is more likely to assign a peak at locations with more biotinylated Pol2. Genes with lower expression may have enough Pol2 to be able to call a peak, but not enough biotin Pol2.

I then investigated how the distance from the insert affected my reads. I reasoned that given that Pol2 is biotinylated at the integrated Tet operon, and the location of the Tet operon is known (chromosome 12, position 1223244) then biotinylated Pol2 will diffuse and reach closer genes before more distant ones. I calculated the coverage normalized Log2 fold ratio of tetracycline induced to not induced samples from each concentration of plasmids infected at all gene promoters (TSS +/- 250 bp) on chromosome 12 using deeptools (Ramírez et al., 2016). This will allow me to calculate the enrichment after tetracycline release. I hypothesised that there would be a negative correlation between Log2 fold ratio and distance from the insert. I then constructed a linear model predicting Log2 fold ratio at gene promoters against genetic distance from the insertion site. In all cases I found no significant interaction (P = 0.50, 0.94, 0.65 for plasmid concentrations high, medium, and low respectively). I then reasoned that three-dimensional distance affecting diffusion is not necessarily the same as one-dimensional distance in base pairs. For example, sequences inside a topologically associating domain (TAD) are in closer three-dimensional proximity than sequences outside a TAD (Pombo and Dillon, 2015), even if the distance in base pairs is longer. TAD sizes are known to vary, as well as to be sensitive to the method chosen to call them (Zufferey et al., 2018), but are generally between 100 kb – 5 MB (Rocha et al., 2015). I took genes that were closer to the insert than 2.5 MB, as these
are likely to be in the same, or a nearby TAD to the insert. This resulted in 94 genes, which I hypothesised would have a higher Log\(_2\) fold ratio than the remaining 2960 genes from the insert. I found no significant difference in the high or low concentrations of plasmid transfected (\(P = 0.89, 0.45\) respectively, two-tailed T test) but a significantly higher Log\(_2\) fold ratio for the closest 94 in medium concentrations of plasmid (\(P = 0.0079\), two-tailed T test) (Figure 31). This tentatively suggests that biotinylated Pol2 behaves as I hypothesised, diffusing away from the Tet operon. This result may also suggest that the system can be optimized further – for example by improving the timings of how long tetracycline is induced for, and how much Avi-RPB1 and TetR-BirA is transfected.

![Figure 31. Boxplot showing the Log\(_2\) fold ratio of promoter coverage for the closest 94 genes to the insert (within 2.5 Mb) and the remaining 2960 genes. Biotin Pol2 appears to be higher for the genes nearest the insert for plasmid 148 ng plasmid transfected (middle), significance is shown with **, representing \(P=<0.01,\) two-tailed \(T\) test)](image)

**Summary**

Overall, I have successfully developed a system where Avi-RPB1 is biotinylated at the Tet operon by the fusion protein TetR-BirA. The system requires all three components to work correctly, demonstrated through western blotting. Pulling down biotinylated Pol2 with streptavidin showed that biotinylated Pol2 behaves like total Pol2 (Figure 28 and 29 A), and that uniquely Tet induced peaks appear even more
similar to total Pol2 ChIP-seq (Figure 30 A). This may suggest that some other biotinylated factors may also bind DNA and are pulled down along with biotin-Pol2. I also find tentative evidence that genes nearby the Tet operon have more biotin-Pol2 reads than those further away after induction, suggesting that biotin-Pol2 diffuses away from the Tet operon after induction (Figure 31). I was surprised however, that I did not see a large amount of biotinylated Pol2 at the Tet operon before induction (Figure 30 C), which could potentially be explained by abortive transcription (Part 2, Chapter 4).

Even though induction with tetracycline appears to show increased release of biotinylated polymerase from the Tet operon (Figure 28 D), it is difficult to measure recycling at HBB. This is for two reasons. Firstly, there appears to be a low number of reads at HBB before and after tetracycline induction. Secondly, there appears to be release of biotinylated Pol2 before tetracycline induction from the Tet operon, demonstrated by the metagene plots. Therefore, from ChIP Sequencing by pulling down biotinylated Pol2 alone, it is impossible to distinguish between recycled (biotinylated) polymerases and those that have previously been released (without initiation) and then later initiated but are still biotinylated. These two reasons mean that ChIP-Seq pulling down biotinylated Pol2 alone cannot be used to measure recycling.

I will discuss in the next chapter that ChIP-qPCR can be used to measure both biotinylated and total Pol2, which can be used to determine recycling. This approach can also help solve the first problem, as qPCR is sensitive to small numbers of reads.
Chapter 2: Developing a model of transcription to make predictions on polymerase recycling

Introduction

As mentioned in the previous chapter, ChIP-qPCR can be used to resolve the problems of ChIP-seq. Firstly, it allows enrichment of a particular area of interest, so areas with low coverage can be studied without the cost of increasing sequencing depth.

The second problem which could be resolved is that of release from the Tet operon without tetracycline induction. I hypothesised that calculating the ratio of biotinylated to total Pol2 could be used as a measure of recycling, and to formalize this hypothesis I constructed a toy mathematical model to try and understand the system. I used this model to see if it would be possible to distinguish between altering recycling or initiation rates at HBB to reflect the potential effect of the point mutation in the poly-adenylation site.

I started by constructing a simple model (Model A, Figure 32 A) where free Pol2 can move between HBB and a control promoter (GAPDH), be transcribed (moving between promoter and gene), and either recycled (back to promoter) or moving back to the free pool of polymerase.

In a nutshell, if Pol2 is completely locally recycled, it will remain at HBB and not appear at GAPDH. If not, it should gradually increase at GAPDH.

Results

I constructed a system of ordinary differential equations (ODEs) to describe this model and implemented it using the package DeSolve v1.28 (Karline Soetaert, Thomas Petzoldt, 2010). This is similar to what was measured with (Mapendano et al., 2010) who measured the occupancy of Pol2 at the HBB promoter compared to the GAPDH promoter.
Figure 32. (A) Model A. Diagram of a toy model of transcription. Here, Free Pol2 initiates transcription at 2 promoters, HBB and GAPDH (a control gene). Pol2 then initiates transcription and can either be recycled back to the promoter or move off the gene. (B) Model B. This is similar to model A but distinguishes between biotinylated Pol2 and total.

\[
\begin{align*}
\frac{d\text{Free Pol2}}{dt} &= k_{\text{on}}^{\text{GAPDH}} \times \text{Free Pol2} + k_{\text{recycle}}^{\text{HBB}} \times \text{Pol2}^{\text{HBB gene}} - k_{\text{transcription}}^{\text{HBB}} \times \text{Pol2}^{\text{HBB promoter}} - k_{\text{initiation}}^{\text{HBB}} \times \text{Free Pol2} - k_{\text{off}}^{\text{HBB}} \times \text{Pol2}^{\text{HBB gene}} \\
\frac{d\text{Pol2 HBB promoter}}{dt} &= k_{\text{initiation}}^{\text{HBB}} \times \text{Free Pol2} + k_{\text{recycle}}^{\text{HBB}} \times \text{Pol2}^{\text{HBB gene}} - k_{\text{transcription}}^{\text{HBB}} \times \text{Pol2}^{\text{HBB promoter}} \\
\frac{d\text{Pol2 GAPDH promoter}}{dt} &= k_{\text{on}}^{\text{GAPDH}} \times \text{Free Pol2} + k_{\text{recycle}}^{\text{GAPDH}} \times \text{Pol2}^{\text{GAPDH gene}} - k_{\text{transcription}}^{\text{GAPDH}} \times \text{Pol2}^{\text{GAPDH promoter}} \\
\frac{d\text{Pol2 HBB gene}}{dt} &= k_{\text{transcription}}^{\text{HBB}} \times \text{Pol2}^{\text{HBB promoter}} - k_{\text{recycle}}^{\text{HBB}} \times \text{Pol2}^{\text{HBB gene}} - k_{\text{off}}^{\text{HBB}} \times \text{Pol2}^{\text{HBB gene}} \\
\frac{d\text{Pol2 GAPDH gene}}{dt} &= k_{\text{transcription}}^{\text{GAPDH}} \times \text{Pol2}^{\text{GAPDH promoter}} - k_{\text{recycle}}^{\text{GAPDH}} \times \text{Pol2}^{\text{GAPDH gene}} - k_{\text{off}}^{\text{GAPDH}} \times \text{Pol2}^{\text{GAPDH gene}}
\end{align*}
\]
\[ \frac{d \text{Free BiotinPol2}}{dt} = k_{\text{off}}^{\text{GAPDH}} \times \text{BiotinPol2}^{\text{GAPDH gene}} + k_{\text{off}}^{\text{HBB}} \times \text{BiotinPol2}^{\text{HBB gene}} - k_{\text{initiation}}^{\text{GAPDH}} \times \text{Free BiotinPol2} - k_{\text{initiation}}^{\text{HBB}} \times \text{Free BiotinPol2} \]

\[ \frac{d \text{BiotinPol2 HBB promoter}}{dt} = k_{\text{on}}^{\text{HBB}} \times \text{Free BiotinPol2} + k_{\text{recycle}}^{\text{HBB}} \times \text{BiotinPol2}^{\text{HBB gene}} - k_{\text{transcription}}^{\text{HBB}} \times \text{BiotinPol2}^{\text{HBB promoter}} \]

\[ \frac{d \text{BiotinPol2 GAPDH promoter}}{dt} = k_{\text{on}}^{\text{GAPDH}} \times \text{Free BiotinPol2} + k_{\text{recycle}}^{\text{GAPDH}} \times \text{BiotinPol2}^{\text{GAPDH gene}} - k_{\text{transcription}}^{\text{GAPDH}} \times \text{BiotinPol2}^{\text{GAPDH promoter}} \]

\[ \frac{d \text{BiotinPol2 HBB gene}}{dt} = k_{\text{transcription}}^{\text{HBB}} \times \text{BiotinPol2}^{\text{HBB promoter}} - k_{\text{recycle}}^{\text{HBB}} \times \text{BiotinPol2}^{\text{HBB gene}} - k_{\text{off}}^{\text{HBB}} \times \text{BiotinPol2}^{\text{HBB gene}} \]

\[ \frac{d \text{BiotinPol2 GAPDH gene}}{dt} = k_{\text{transcription}}^{\text{GAPDH}} \times \text{BiotinPol2}^{\text{GAPDH promoter}} - k_{\text{recycle}}^{\text{GAPDH}} \times \text{BiotinPol2}^{\text{GAPDH gene}} - k_{\text{off}}^{\text{GAPDH}} \times \text{BiotinPol2}^{\text{GAPDH gene}} \]

Equation 2. Ordinary differential equations showing the rate equations governing model B. The rate equations for model A are the same as the first 5 rate equations shown here.
Whilst this was a toy model, I wanted the relative rates to be somewhat realistic. I determined the average Pol2 coverage for a protein coding gene inside the promoter region and the gene body using ChIP-Seq datasets from (McHaourab et al., 2018). The promoter region was defined as (TSS +/- 250 bp), and the gene body was defined as (TSS+250 bp to TES -250 bp). I found that the median Pol2 coverage at the promoter was 1.8 times higher than the median coverage in the gene body. I started by setting every parameter in Model A to the same value (0.1), with the starting condition that all Pol2 is unbound (Figure 33 A). At equilibrium, this model yields double the Pol2 coverage at the promoter, compared to the gene body. Setting $k_{recycle}$ at both HBB and GAPDH to 0 required raising $k_{off}$ to 0.2 at both HBB and GAPDH to maintain this 2:1 ratio (Figure 33 B).

Using this model (model A), I generated 2 simulations, which I used to reflect the experiments from (Mapendano et al., 2010). In each of these simulations, the starting conditions for Pol2 was, 0.5 Pol2$^{\text{Free}}$, 0.2 Pol2$^{\text{HBB promoter}}$, 0.2 Pol2$^{\text{GAPDH promoter}}$ and 0.1 Pol2$^{\text{GAPDH gene}}$. I reasoned this reflected the 2:1 ratio of Pol2 at the promoter:gene for GAPDH, but as HBB had not been induced, there is no Pol2 at this gene.

![Diagram](image)
In the first simulation, I investigated the effects of decreasing the initiation rate at HBB, which could potentially reflect the pA-cell line. I reduced $k_{\text{initiation}}^{\text{HBB}}$ to 0.02 and then calculated the ratio of Pol2 at the HBB promoter compared to the GAPDH promoter, as this is what was measured in (Mapendano et al., 2010). I then compared this ratio to maintaining the $k_{\text{initiation}}^{\text{HBB}}$ to 0.1, which I used to reflect the WT cell line (Figure 34 A). I also investigated the effects of removing the ability of HBB to be recycled, which could also reflect the pA-cell line. Here, I reduced $k_{\text{recycle}}^{\text{HBB}}$ to 0, to completely disrupt recycling, and I calculated the ratio of Pol2 at the HBB promoter compared to the GAPDH promoter. Again, I compared this to maintaining $k_{\text{recycle}}^{\text{HBB}}$ at 0.1, reflecting the WT cell line (Figure 34 B). Whilst the exact model outputs are dependent on the conditions chosen, it is clear that both of these outcomes are qualitatively consistent with the previously published reports (Mapendano et al., 2010), and therefore using ChIP-qPCR on Pol2 alone, cannot differentiate between differences in initiation and differences in recycling.

**Figure 34.** Figure demonstrating that model A cannot distinguish between disruption of initiation or recycling in the pA-cell line. Note the order of the lines match each other, regardless of if recycling or initiation is altered. (A) Model simulation where all rates are kept at 0.1 for the blue curve, representing the WT line. The orange curve represents the pA-line, where $k_{\text{initiation}}^{\text{HBB}}$ is reduced to 0.02. (B) Model simulation where all rates are kept at 0.1 in the blue curve, again representing the WT line. In the orange curve $k_{\text{recycle}}^{\text{HBB}}$ is reduced to 0, with all other rates kept at 0.1, this represents the pA-line.
I also constructed a second model (model B, Figure 32 B, Equation 2), which differentiates between biotinylated Pol2 and total Pol2. Again, I constructed a system of ODEs and implemented it using DeSolve v1.28 (Karline Soetaert, Thomas Petzoldt, 2010). In this model I have made some assumptions based on prior knowledge of the system I developed. Firstly, I have assumed that most Pol2 is not biotinylated, and so the distribution of biotinylated Pol2 does not affect the distribution of total Pol2. The other assumption I have made is that biotinylated Pol2 and Pol2 behave in the same way, so their reaction rates are the same. I then chose some starting conditions for my model.

I started by constructing a model describing an idealised version of my system. Here, there is no total-Pol2 at HBB promoter (biotinylated Pol2 is a small proportion of total Pol2) or on the HBB gene as it blocked from transcribing by TetR-BirA. Total Pol2 is distributed as freely diffusing Pol2 (0.7), on the GAPDH promoter (0.2), or on the GAPDH gene (0.1). As Pol2 is biotinylated at the Tet operon, I started by assuming that most biotinylated Pol2 starts at the HBB promoter (0.7) but is also present at GAPDH promoter (0.1), GAPDH gene (0.1) and is freely diffusing (0.1). I note that ChIP-Seq experiments (Figure 29 B-C) show a low proportion of reads at the HBB promoter, and later simulations investigate this.

With these starting conditions, this model can distinguish between recycling and initiation differences. This is through looking at the ratio of biotinylated Pol2 to total Pol2.

\[
\text{Ratio} = \frac{\frac{\text{BiotinPol2}^{HBB \text{ promoter}}}{\text{Pol2}^{HBB \text{ promoter}}}}{\frac{\text{BiotinPol2}^{GAPDH \text{ promoter}}}{\text{Pol2}^{GAPDH \text{ promoter}}}}
\]

*Equation 3. Ratio of biotinylated Pol2 to total Pol2 at HBB, normalised by GAPDH.*

With these starting conditions, I created 2 simulations with the same rates as presented in Figure 34. For the first simulation I altered initiation at the HBB
promoter which could reflect the pA- and WT cell lines (k-initiation$^{HBB}$ is 0.02 and 0.1 respectively), and for the other I altered recycling at the HBB locus which could also reflect the pA- and WT cell lines (k-recycle$^{HBB}$ is 0 and 0.1 respectively). Here, I found that it is possible to distinguish between these two cases (Figure 35). If increased initiation in the WT line caused the increase in Pol2 at the HBB promoter compared to the pA- line, then my model predicts that the ratio of biotinylated to total Pol2 would be higher in the pA- line (Figure 35 A). If, however, increased recycling in the WT line caused this effect, then the ratio would be higher in the WT line (Figure 35 B). This means that ChIP-qPCR could be used to measure if polymerase recycling or initiation is altered by mutating the poly-adenylation signal.

Figure 35. Model simulations of model B, which demonstrate that calculating the ratio of Biotin-Pol2 to total Pol2 at HBB compared to GAPDH can distinguish between disrupted recycling and initiation in the pA- line. Note the order of the lines switch around, depending on whether recycling or initiation is altered. In both models, the blue curve represents the WT cell lines, and the orange curve represents the pA- line. (A) Model simulation where all rates are kept at 0.1 in the blue curve, in the orange curve $k$-initiation$^{HBB}$ is reduced to 0.02. (B) Model simulation where all rates are kept at 0.1 in the blue curve, in the orange curve $k$-recycle$^{HBB}$ is reduced to 0.

Potentially, different rates in my model could result in different results which could question conclusions derived from this model. I therefore conducted a sensitivity analysis to investigate the parameter space of my model. I held the starting conditions constant as described above, and I held the rates at GAPDH to equal the
rates at HBB – excluding either initiation or recycling which I used to reflect the pA-line.

Firstly, I investigated the parameter space when altering recycling. I explored changing each rate ($k_{\text{initiation}}$, $k_{\text{transcription}}$, $k_{\text{off}}$ and $k_{\text{recycle}}$) through 5 orders of magnitude (0.01, 0.1, 1, 10, 100). I then changed how much the point mutation could affect the recycling at HBB by multiplying $k_{\text{recycle}}^{\text{HBB}}$ by 0, 0.2, 0.4, 0.6 and 0.8 to replicate potential disrupted recycling in the pA-line. In total this resulted in evaluating 625 sets of differential equations for each cell line. In all cases, the line with higher $k_{\text{recycle}}^{\text{HBB}}$ had a higher, or equal biotin-Pol2 to total Pol2 ratio at the promoter (HBB/GAPDH) than the line with reduced recycling.

Secondly, I took the same approach – but instead of changing recycling to replicate the pA-line, I altered $k_{\text{initiation}}^{\text{HBB}}$ by multiplying this by 0.01, 0.2, 0.4, 0.6 and 0.8 to replicate the potential of disrupted initiation on the pA-line. Again, this resulted in evaluating 625 sets of differential equations for each cell line, and in all cases the lines with higher $k_{\text{initiation}}^{\text{HBB}}$ had a lower, or equal biotin-Pol2 to total Pol2 ratio at the promoter (HBB/GAPDH).

I then decided to investigate altering the rates at GAPDH, whilst holding the rates at HBB constant. I altered $k_{\text{initiation}}^{\text{GAPDH}}$, $k_{\text{transcription}}^{\text{GAPDH}}$, $k_{\text{off}}^{\text{GAPDH}}$ and $k_{\text{recycle}}^{\text{GAPDH}}$ through 5 orders of magnitude (0.01, 0.1, 1, 10, 100). All rates at HBB were kept at 0.1, apart from $k_{\text{recycle}}^{\text{HBB}}$ reflecting the pA-line which I reduced to 0, reflecting possible disrupted recycling in this line. Here I also found that, in all cases, the lines with higher $k_{\text{recycle}}^{\text{HBB}}$ had a higher or equal ratio than those with reduced $k_{\text{recycle}}^{\text{HBB}}$.

I then took the same approach as above, and altered the rates at GAPDH through 5 orders of magnitude. I maintained rates at HBB equal to 0.1, excluding $k_{\text{initiation}}^{\text{HBB}}$ which I reduced to 0.02 to possibly reflect the pA-line. Here, again in all cases, the lines with higher $k_{\text{initiation}}^{\text{HBB}}$ had a lower, or equal biotin-Pol2 to total Pol2 ratio at the promoter (HBB/GAPDH).
My ChIP-Seq experiments from Part 2, Chapter 1 (Figure 29 B-C) suggest that there is a low proportion of biotinylated Pol2 at the HBB promoter, which is unlike the idealised version of my system I have modelled. I compared the coverage at the HBB promoter to the GAPDH promoter and found that there was approximately 1.5 times the coverage at GAPDH compared to HBB before polymerase induction. I therefore adjusted the starting conditions for my model. I changed the starting proportion of biotinylated polymerases to: HBB promoter (0.1), GAPDH promoter (0.2), GAPDH gene (0.1) and freely diffusing (0.6) to reflect this. I kept the total polymerase concentrations, and rate constants the same as Figure 35. I found qualitatively the same results as before and found that it is possible to distinguish between increased recycling or initiation from the WT cell line (Figure 36), though the difference between the WT and the pA- line decreased.

Figure 36. Model simulations of model B, as Figure 35, but with more realistic starting conditions. This shows that it is possible to distinguish between increased initiation or increased recycling in the WT line. Note the order of the lines switch around, depending on whether recycling or initiation is altered. In both models, the blue curve represents the WT cell lines, and the orange curve represents the pA- line. (A) Model simulation where all rates are kept at 0.1 in the blue curve, in the orange curve \( k_{\text{initiation}}^{\text{HBB}} \) is reduced to 0.02. (B) Model simulation where all rates are kept at 0.1 in the blue curve, in the orange curve \( k_{\text{recycle}}^{\text{HBB}} \) is reduced to 0.
From this exploration of parameter space, I have shown that the results are insensitive to the particular choice of rate parameters. In all cases studied, if the WT line has higher recycling at HBB than the pA- line, then the ratio of biotin-Pol2 to total Pol2 at the promoters (HBB/GAPDH) is higher. If, however, the WT line has higher initiation than the pA- line, then the ratio is lower. Therefore, this suggests that ChIP-qPCR could be used, measuring the biotin-Pol2 to total Pol2 ratio at the promoter (HBB/GAPDH) to discover if polymerase recycling or initiation is altered by mutating the poly-adenylation signal.

I then tested some model simulations to determine if differences in the starting conditions between the two cell lines could influence the results. Here, I generated two simulations, one where the amount of biotinylated Pol2 on the HBB promoter was halved (0.35, originally 0.7, Figure 37 A), and another where amount of total Pol2 on the HBB promoter was increased to 0.3 (originally 0, Figure 37 B), with no changes in rates. I compared these results to the previous simulations (Figures 35 and 36) and found that altering the starting conditions could affect the results from ChIP-qPCR. This is because having differing initial biotin-Pol2 or total Pol2 on the HBB promoter in one line than the other could give qualitatively similar results to changes in recycling or initiation.
The starting conditions are unlikely to be different between the two cell lines, as they are clones of each other, with the only difference being the point mutation on the poly-adenylation signal. This means transfection efficiency and the total amount of Pol2 is unlikely to be different between the cell lines. The total amount of polymerase at the promoter before Tet induction is also not significantly different between the two cell lines (Mapendano et al., 2010). Furthermore, without induction, there is very minimal leaky transcription in either of the two lines (Cavallaro et al., 2021), making the point mutation on the poly-adenylation signal unlikely to affect the accumulation of biotinylated Pol2 on the promoter. Whilst the starting conditions are unlikely to be different between the two cell lines, this can be measured, and I decided to measure both the total amount of Pol2, and biotinylated Pol2 in the two cell lines to see if the starting conditions were the same.
Proteomics

To further interrogate our system, I wanted to quantify the total amount of Pol2, and biotinylated Pol2, to see if they are different between the two cell lines in my model.

I started by quantifying RPB1 (large subunit of Pol2). To do this I followed the same starting protocol as for the ChIP-qPCR experiments, but I loaded 60 μg of total protein onto a commercial RPB1 Elisa kit and followed the manufacturer’s recommended protocol. I conducted 3 independent replicates of both cell lines after transfection.

The standard curve appeared robust with an $R^2$ value of 0.999, which I used to quantify the concentration of total RPB1 (Figure 38 A). I found that there was no significant difference between the two cell lines (WT & pA-) ($p=0.42$, two tailed t-test) in the total concentration of RPB1 (Figure 38 B). As there was no significant difference between the two cell lines, the assumption that my model makes is not violated.

Figure 38. (A) Standard curve of ELISA assay ($R^2 = 0.999$) (B) Boxplot showing RPB1 concentration (ng/L) when 60 μg of total protein was loaded for each sample. There is no significant difference between the WT and the pA- cell lines ($P=0.42$, two tailed t-test).
I also wanted to discover if the abundance of biotinylated Avi-RPB1 is the same between the two cell lines. I decided to use Mass spectrometry to quantify the relative concentrations of biotinylated Avi-RPB1.

To analyse the starting conditions of biotinylated Pol2 between the two conditions I first needed to purify the proteins in my sample to enrich for biotinylated Pol2. This is because unbiotinylated Pol2 will also be present in my sample, meaning that distinguishing between biotinylated and non-biotinylated Pol2 would be difficult, as biotinylation only occurs on the Avi-tag. I then planned to measure if there is a difference in biotinylated Avi-RPB1, relative to all other biotinylated proteins.

Initially I decided to purify biotinylated Pol2 using a soft-link avidin resin, which allows binding of biotin bound proteins and release with the addition of excess biotin. I then planned to purify the eluted proteins with Tris-Tricine SDS-PAGE, and follow an in-gel digestion protocol to prepare the samples for mass spectrometry.

Unfortunately, after multiple attempts I did not see any proteins purified at all using a soft link avidin resin. I therefore decided to use the same streptavidin beads that I previously used for ChIP-Seq experiments as I know these beads successfully pull down biotinylated Pol2. However, streptavidin is known to bind biotin very tightly, and is one of the strongest non covalent interactions found in nature (Chivers et al., 2011), making it difficult to break this interaction.

This interaction has been shown to be broken by heating to above 70°C for a few minutes (Holmberg et al., 2005), and so I decided to pull down biotinylated Pol2 with streptavidin and elute with heating. Unfortunately, after multiple attempts I still was unable to purify any proteins.

I therefore decided to pull down biotinylated proteins using streptavidin beads and do an on-bead digestion to prepare the samples for mass spectrometry, following a similar protocol to bioID (Roux et al., 2018). This has the disadvantage of also
digesting streptavidin from the beads, resulting in them appearing in the mass spectrometry output.

To do this, I followed my ChIP-qPCR experiments as much as possible, but instead of eluting DNA off the streptavidin beads, I digested the proteins off the streptavidin beads using a tryptic digest.

I started by conducting a pilot mass spectrometry experiment, looking only at HBB WT which was either transfected, or not transfected with the two fusion proteins (Avi-RPB1 and BirA-TetR). In total, 116 proteins in 97 clusters were identified (Figure 39 C).

<table>
<thead>
<tr>
<th>Protein</th>
<th>Transfected Abundance [Normalized total spectra]</th>
<th>Negative control Abundance [Normalized total spectra]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CON__P02769 (bovine albumin precursor)</td>
<td>410</td>
<td>358</td>
</tr>
<tr>
<td>Keratin, Type 2</td>
<td>110</td>
<td>107</td>
</tr>
<tr>
<td>Avi-RPB1</td>
<td>70</td>
<td>0</td>
</tr>
<tr>
<td>Con_Streptavidin</td>
<td>67</td>
<td>34</td>
</tr>
<tr>
<td>Serum albumin</td>
<td>56</td>
<td>52</td>
</tr>
<tr>
<td>Keratin, Type 1</td>
<td>41</td>
<td>46</td>
</tr>
</tbody>
</table>

Figure 39 A) The top 6 identified proteins sorted by log₂ abundance in the transfected sample. B) A scatter plot plotting the abundance of all identified proteins with the ideal 45° line drawn, the circled protein shows Avi-RPB1 which is identified in the transfected sample, but not in the negative control. C) Venn Diagram showing the high degree of overlapping proteins identified between the two samples (P = 2.7e-188, hypergeometric test).

I found that the transfected and non-transfected samples replicated each other well with a high degree of overlap of individual proteins detected (P = 2.7e-188, hyper
geometric test, total number of protein coding genes = 20589) as well as their normalized abundance.

Encouragingly, I found proteins we expected in the mass spectrometry output (Figure 39 A). Bovine albumin and streptavidin are present, which are both components of the streptavidin beads. Albumin is also a component of the cell culture serum. I also found common mass spectrometry contaminants such as Keratins which have previously been discussed in detail (Hodge et al., 2013).

I also found known biotinylated carboxylases in both samples (pyruvate carboxylase, methylcrotonoyl-CoA carboxylase subunit alpha, acetyl-CoA carboxylase 1A, and propionyl-CoA carboxylase alpha chain) (Roux et al., 2018). This is also expected as these are likely to be pulled down by the streptavidin coated beads. I also found histones present in the output in low abundances, potentially supporting my suggestions regarding regions of the genome that are bound by other biotinylated proteins, from Part 2, Chapter 1.

From this experiment, I had two major results. The first result was that we confirmed that streptavidin beads do indeed pull-down biotinylated Avi-RPB1, which confirm the results from our western blots. Our second major result was that the highest abundance protein in the transfected sample which is not a known contaminant for mass spectrometry is Avi-RPB1 (Figure 39 A). This is absent in the non-transfected sample (Figure 39 B). The fact that Avi-RPB1 is the highest abundance protein that is not a known contaminant also provides evidence for the specificity of our approach to use streptavidin to pulldown for ChIP-Seq and ChIP-qPCR.

I then decided to continue with a full-scale experiment to measure the abundance of biotinylated RPB1 between the two cell lines (WT & pA). I conducted 3 independent replicate transfections for each cell line as well as negative controls (not transfected) for each cell line respectively.
WT sample 1 was excluded from the analysis as only 3 proteins/clusters were identified, P02769 (cluster of albumins), P00761 (Trypsin) and ENSBTAP00000031900 (Bovine Ceruloplasmin). Ceruloplasmin is a copper carrying protein present in blood and is therefore likely to be a contaminant from the FBS. It is therefore likely that this sample was not successfully prepared, and as such was removed from the analysis. I found that there was no significant difference between both cell lines in the amount of biotinylated Avi-RPB1 (P=0.46, two tailed T-test). As there was no significant difference between the two cell lines, the assumption that my model makes is not violated, and provides evidence to using ChIP-qPCR to measure polymerase recycling (Figure 40).

Figure 40. A scatter graph showing the abundance of biotinylated proteins (Normalized total spectra) on the logarithmic scale. The dotted line shows the ideal 45° line between the cell lines. The marked dot shows Avi-RPB1, which is not significantly differentially expressed between the pA- and the WT cell line (P=0.46, two tailed t-test).
Summary

In this section of my thesis, I developed a toy model consisting of a system of differential equations to model transcription. I used this to investigate different possible scenarios by conducting simulations altering initiation and recycling rates to reflect the two cell lines. I discovered that measuring the ratio of biotinylated Pol2 to total Pol2 at the HBB promoter (compared to a control gene’s promoter) could determine if the local recycling of Pol2 was different between two cell lines. I found that differing starting conditions between the cell lines could affect my results. I therefore conducted an ELISA and a mass spectrometry experiment and found that there is no significant difference in total Pol2 and biotinylated Pol2 between the two cell lines before tetracycline induction.
Introduction

The results from both my simplified model of transcription and the knowledge that biotinylated and non-biotinylated Pol2 was not present at different abundances between the two cell lines (Part 2, Chapter 2), suggest that ChIP-qPCR can be used to measure polymerase recycling. I measured biotinylated Pol2 and total Pol2 occupancy at the HBB promoter and the promoter of a control gene, GAPDH in two cell lines – HBB WT and HBB pA- (Figure 41 shows a schematic of the experimental protocol). My hypothesis is that differences in these ratios between the cell lines will indicate if there is a difference in recycling or initiation (Part 2, Chapter 2).

Results

I used modified qPCR primers from (Mapenando et al., 2010) to measure polymerase occupancy at the promoters. These were modified to increase the melting temperature to 60°C as recommended by the manual for the qPCR master mix (PowerUp SYBR Green Master Mix, Thermo Fisher). Unfortunately, the antibody
against Pol2 that was used in Mapendano et al. (2010) had been discontinued so I used one raised against Pol2’s CTD (SC:47701). This is because the biotinylation occurs on the N terminal domain, at the Avitag, and is unlikely to affect this antibody.

I then calculated my primer efficiencies which were both acceptable and similar to the results from (Mapendano et al., 2010). These were calculated to be 1.09 and 1.07 for HBB and GAPDH, respectively. This was calculated using the default method using the “qpcR” package (Spiess, 2018). I note it is surprising to have PCR efficiencies over 1, however these are known to sometimes occur and are between the rule of thumb of acceptability ranging from 0.9-1.10.

I conducted a ChIP-qPCR experiment of DNA (100 pg for each sample) pulled down with the Pol2 antibody or streptavidin. I used the HBB and GAPDH (a control gene) primers to measure the ratio of Biotin-Pol2:Total Pol2 enrichment for both cell lines (Equation 4).

\[
\text{Enrichment}_{\text{Biotin-pol2}} = 2^{-\Delta ct} \\
\text{where } \Delta ct = HBB_{ct(\text{Biotin-pol2})} - GAPDH_{ct(\text{Biotin-pol2})}
\]

\[
\text{Enrichment}_{\text{Total-pol2}} = 2^{-\Delta ct} \\
\text{where } \Delta ct = HBB_{ct(\text{Total-pol2})} - GAPDH_{ct(\text{Total-pol2})}
\]

\[
\text{Biotin Pol2 to total Pol2 ratio} = \frac{\text{Enrichment}_{\text{Biotin-pol2}}}{\text{Enrichment}_{\text{Total-pol2}}}
\]

*Equation 4. Calculation of the biotinylated Pol2 to total Pol2 ratio.*

Melt curve analysis was undertaken which showed one clear peak in all samples, and no peak in the no template controls (NTC). This single peak indicates only 1 PCR product, which shows specific qPCR. Figure 42 demonstrates this from a typical experiment.
The cycle threshold (ct) was calculated using the default parameters of the qpcR package (Spiess, 2018). 5 biological repeats were undertaken, and enrichment of biotin-Pol2 and total-Pol2 were calculated. Enrichments which lay outside of Tukey’s Fence (1.5X IQR) were discarded as these were judged to have failed. 5 samples were excluded from Biotin Pol2 enrichments, and 6 from total Pol2 enrichments, from a total of 160 qPCR experiments. The ratio was then calculated as described earlier.

I found that biotin-Pol2 enrichment was significantly higher in the WT line compared to the pA- line after tetracycline induction (Figure 43 A, two-tailed t-tests), whereas I found no significant differences between total pol2 enrichment in the two cell lines (Figure 43 B, two-tailed t-tests). I then calculated the ratio of the two enrichments, which was significantly higher in the WT line compared to the pA- line (Figure 43 C, two-tailed t-tests).

Figure 42. ChIP-qPCR melt curves of all samples (one representative biological replicate). S represents streptavidin pulldown (measuring biotinylated Pol2), Pol represents Pol2 antibody pulldown. The black line represents fluorescence (F) vs temperature (T) and the red line is $dF/dT$ against temperature. The single peak on the red line represents a single PCR product indicating the specificity of the qPCR, for all samples, excluding the Non-template control (NTC).
Interestingly, my results do not exactly match with Mapendano et al. (2010). I do not see any significant difference between total-Pol2 enrichment at HBB compared to GAPDH, whereas they find significantly higher total-Pol2 enrichment. Broadly, there are two major differences between mine and their protocol. They use a different antibody to mine, and they load more DNA for qPCR than my protocol. The antibody used in Mapendano et al. (2010) (sc-9001) was a rabbit polyclonal to Pol2. This was discontinued by the supplier and I used the suppliers recommended replacement (sc-47701), a monoclonal antibody against the Pol2 CTD. These antibodies are unlikely to behave differently to each other as they have both been extensively used for total Pol2 pulldowns.

Mapendano et al. (2010) also load approximately 1ng of DNA for qPCR (estimate from supplementary information) whereas I add 100pg of DNA. The reason I add less DNA
is that I have a low yield of DNA after streptavidin pulldown, likely due to the small numbers of Polymerases biotinylated. I use the same DNA concentration for total Pol2 and biotinylated Pol2 pulldowns for qPCR which results in approximately tenfold reduced DNA loading. It is possible that the difference between my results and Mapendano et al. (2010) is due to increased technical variation caused by low amounts of DNA added, which is supported by the wide variation on the boxplots.

Furthermore, my models predicted the ratio to level out at 1, whereas I see the ratio falling below 1. This may be explained by streptavidin pulling down some factors other than biotinylated Pol2, which could be present at the GAPDH promoter, for example biotinylated histones (Part 2, Chapter 2).

Summary

Overall, even with the wide variation, the results match the model predictions, and my hypothesis. There is a higher ratio of biotinylated Pol2 to total Pol2 at HBB compared to GAPDH in the WT cell line compared to the pA- cell line. This provides further evidence to previously published research from our laboratory (Cavallaro et al., 2021), that the local recycling of RNA Pol2 at HBB is higher in WT than the pA- cell lines.

This evidence is orthogonal to our previous work (Cavallaro et al., 2021), which was based on fitting models to HBB mRNA levels in both cell lines. Here, I use a more direct method to measure Pol2 which was originally at the start of the gene (near to Tet operon) and shows that this labelled Pol2 is maintained at HBB more in the WT cell line than the pA- cell line. This provides evidence that 3’-5’ crosstalk through the mechanism of polymerase recycling contributes to the increased transcriptional noise found at the HBB locus.
Chapter 4: Measuring abortive transcription through a model gene using labelled Pol2 ChIP-Seq

Introduction

One question raised by the previous chapter is why biotinylated Pol2 is present on other genes before tetracycline induction (e.g. Figure 28 A-C), as TetR-BirA is expected to block transcription, stop polymerase progression and hold biotinylated Pol2 at the HBB promoter. This presence at other genes is unlikely to be due to spurious biotinylation of Pol2, as biotinylation only occurs at the Tet operon. I have showed this through western blotting using streptavidin-HRP on cell lines, transfected with Avi-RPB1 and TetR-BirA which contain the Tet operon, and those which do not contain the Tet operon (Figure 22 A). I only found a Pol2 band in the cell line that contained the Tet operon.

Furthermore, the system I developed requires all components to be present to biotinylate Pol2; missing any component (Avi-RPB1, TetR-BirA) results in no biotinylation of Pol2, shown through western blotting (Figure 21, 22 B), again demonstrating that the system is specific.

One possibility could be that the TetR-BirA fusion protein is not repressing the expression of HBB. Here, Pol2 could be biotinylated but not repressed by TetR-BirA thus allowing transcription to progress. After termination Pol2 could then spread to other genes which could be seen in the metagene plots.

However, this possibility is unlikely for a few reasons. One reason is that our laboratory have found that there is very minimal leaky transcription of HBB in the HEK293 cell lines that I have used in this study (Cavallaro et al., 2021). Though I note this was measured using the native TetR, and not my fusion protein, and so potentially the fusion protein could cause leaky transcription. The second reason is
that TetR fusion proteins have been made before, for example TetR-GFP (Dublanche et al., 2006), which have also blocked transcription effectively. GFP (28kDa) is of similar size to BirA (36kDa), and this GFP fusion does not result in leaky transcription.

Unfortunately, testing my own TetR-BirA fusion for intact repression is difficult, as the cell lines I have been using also express the (original) Tet repressor. Whilst this is not a problem for the previous work (as all 3 components need to be present for Pol2 to be biotinylated), measuring transcription is different. Not every cell will have successfully transfected the plasmid encoding TetR-BirA, and therefore measuring transcription at HBB will be compounded by a combination of cells expressing only TetR, and cells expressing TetR and TetR-BirA. I would be unable to ensure that TetR-BirA is at the promoter for every cell, and therefore be unable to determine if transcription is blocked by TetR-BirA by measuring HBB expression before and after induction.

One method to address this problem would be to clone a plasmid containing two components: The HBB gene containing the Tet operon and TetR-BirA. This plasmid would be transfected into a human cell line which does not express HBB. Then, HBB expression would be measured with RT-PCR before and after Tetracycline induction. Every cell line which contains HBB with the Tet operon also expresses TetR-BirA, and therefore this could be used to measure how effectively TetR-BirA blocks HBB expression. However, a possible confounder could be timings; HBB will be transcribed upon transfection even if the TetR-BirA fusion protein is effective, as it will take time for TetR-BirA to be translated. Although this can be mitigated, as the half-life for HBB is known (Cavallaro et al., 2021) and western blots can be used to find how long it takes for TetR-BirA to be translated. This means after waiting until the original HBB mRNA is degraded, this system could be used to test how effectively TetR-BirA blocks transcription. Additionally, different timings of transfection will further complicate matters. Therefore, a superior solution would involve these genes being stably inserted into a cell line, which could be grown to equilibrium before testing if HBB is expressed before and after Tetracycline induction.
A second, more appealing explanation could be due to abortive transcription. This term describes a process where during transcription, Pol2 unbinds from the DNA strand and truncated RNA strands are released (See Introduction). Abortive transcription is known to be most common at early stages of transcription, where less than 10% of transcripts can reach productive elongation (Steurer et al., 2018). This may explain why I see reads of biotinylated Pol2 at other genes’ promoters before tetracycline induction (Figure 28 A-C). Here, polymerases may be biotinylated upon transcription initiation and subsequently abort without progressing through the Tet operon.

Abortive transcription can also occur during productive elongation. Productive elongation is a processive process meaning that Pol2 continues to add additional nucleotides without releasing its substrate (the template strand). To add the last nucleotide to an mRNA all the other nucleotides must have been added successfully. If, at any point Pol2 fails to add a nucleotide, then elongation has failed, and the resources required to make a transcript has been wasted. Evolutionarily, it would be beneficial if a transcript is going to be aborted, that this abortion should happen early in transcription, so less resources are wasted on generating a truncated, non-functional mRNA. Following this, these truncated mRNAs could be translated to truncated proteins, which again could waste valuable resources, and also could have deleterious consequences (e.g. Ma et al., 2018).

Abortive transcription is likely to have a larger effect on long genes than short genes. If each individual base has an equal likelihood of polymerases aborting, then overall longer genes will have more total abortions – and much more wasted energy producing non-functional RNAs. This is likely to have deleterious fitness consequences, especially for long genes. For example, DMD is approximately 2.3 Mb in length and requires around 16 hours to be transcribed (Tennyson et al., 1995). This appears hard to tally with average residence times of Pol2 held in productive elongation of 60 minutes (Steurer et al., 2018), measured through treatment with Cordycepin, a nucleoside analogue which inhibits transcript elongation. This residence time is 16 times lower than the residence time needed to complete
transcription of DMD. Therefore, there may be a strong evolutionary pressure to reduce abortive transcription rates as Pol2 progresses through a gene.

In a similar conceptual manner, though in a much different system, abortion of DNA polymerases in assays has been extensively described and modelled before (Archer et al., 2016; Davies et al., 2021; Dyer et al., 2019), in these cases explaining and correcting for biases in RNA-Seq library preparation. These models all assume that each individual base has an equal likelihood of polymerases aborting and demonstrate that this can result in the underestimation of long genes’ expression as these have more chances for an abortive event if this bias is not accounted for.

It would therefore be interesting to see if different genic regions have different abortive transcription rates, and this question may potentially be answered with my system. If I take multiple versions of a model gene and insert the Tet operon in a different location for each version of the model gene, Avi-RPB1 will be biotinylated and blocked from progressing at these locations. Then, after tetracycline induction biotinylated Pol2 will be allowed to progress.

Relative abortive transcription rates could then be measured by comparing the ratio of ChIP-Seq read of biotinylated Pol2 at other gene promoters without tetracycline induction and with tetracycline induction. If abortive transcription is high, then read counts at other genes’ promoters without tetracycline induction would be high in comparison to ones with tetracycline induction. This is because high abortive transcription will result in polymerases being biotinylated at the Tet operon, but then leaving and spreading to other genes as they abort.

In this approach, I have assumed that the TetR-BirA fusion protein does work as expected – and blocks polymerases from progressing through the gene. Whilst this is likely to be true for reasons mentioned above, our laboratory is in the process of testing this. I have also assumed that this blockage of polymerase doesn’t change abortive transcription dynamics, a similar assumption to previous experiments using drugs to block entry into different stages of transcription (Steurer et al., 2018).
Results

I used Gibson assembly (Gibson et al., 2009) to generate plasmids containing a trial gene with the Tet operon inserted into different genic regions. I downloaded the HBB sequence from NCBI, and in silico inserted the Tet operon (2 copies) into 4 genic regions, the 5’ UTR, Intron 1, Exon 3 and the 3’UTR. I designed gblocks to construct these fragments and simulated a Gibson assembly with Snapgene (Figure 44 A-D).

Figure 44. 4 genes constructed with Gibson assembly into an expression vector. (A) HBB with the Tet operon in the 5’UTR, (B) HBB with the Tet operon in intron 1, (C) HBB with the Tet operon inside exon 3 and (D) HBB with the Tet operon in the 3’UTR
I generated overlapping fragments of the gblocks through PCR and linearized the expression vector (pcDNA3, (Sanjabi et al., 2005)) through a double digest using HindIII and EcoR1. I then assembled these fragments using Gibson Assembly® Master Mix (NEB) according to the manufacturer’s instructions and transformed these into competent *E. coli*. Colony PCR was used to screen for successful assemblies which were then confirmed with Sanger sequencing.

I then transfected four 100mm dishes of HeLa cells, with 1.48µg of each plasmid encoding: Avi-RPB1, TetR-BirA and one of HBB with the Tet operon in different locations.

I followed the same optimised protocol as for previous ChIP-Seq experiments, and the day after transfection, I split each dish containing transfected Hela cells into two. The next day I induced HBB expression with 250ng/ml tetracycline for 24 hours for each sample, leaving the other without induction.

I prepared ChIP-Seq libraries using a streptavidin pulldown for all of these and sequenced approximately 6.6 million reads for each sample. After quality control and trimming, these reads were aligned to custom reference genomes (HBB masked hg38, with each plasmid transfected treated as a separate chromosome – including the plasmid containing the Tet operon in the correct location) with bowtie2.

I constructed metagene plots comparing the Log\(_2\)(ratio) of reads for 24 hours post induction to before induction (Figure 45) and expected these to appear similar to the metagene plots in (Fig. 28 D) with a peak at the TSS after polymerase induction. I did expect the signal to noise ratio to be worse in these samples, as this system requires three plasmids to be transfected in the same cell rather than two, which if the chance for each plasmid to be successfully transfected is equivalent and independent will result in the number of cells being transfected with all 3 to be exponentially less likely (Chance\(^3\) rather than Chance\(^2\)).
Whilst these metagene plots do not at first glance appear like the ones from figure 28D, it is interesting to note that for all samples (apart from the sample transfected with the Tet operon at the 5’ end of a gene) I see a sharp rise, at or just after the TSS. Interestingly, this rise appears to be maintained throughout the metagene body, and no clear peak is seen at the TSS, unlike figure 28D.

I presumed this was likely due to the decreased signal to noise from this experiment, compared to the lines containing a genomically integrated Tet operon, along with potentially reduced transfection efficiency or a less efficient pulldown. I decided to repeat the experiment, by preparing new samples and sequencing with a much higher read depth, approximately 12 million reads per sample, to hopefully overcome this decreased signal to noise ratio.
I constructed metagene plots as before, and I found that this replicate performed much more like I expected (Figure 46 A-D). However, the sample containing the exon 3 construct (Figure 46 C) appeared to have the same problems from the first experiment, with no clear peak seen at the TSS. I therefore excluded this sample from quantifying abortive transcription rates.

I then decided to measure relative rates of abortive transcription across a gene. As described above, I reasoned that if the abortive transcription rate is higher, then the difference between 24 hours induction and without induction at other genes’ promoters will be lower. This is because (without induction) Avi-RPB1 is biotinylated at the Tet operon, after abortive transcription the biotinylated Pol2 will unbind from the DNA and bind to other genes’ promoters. This will occur more with higher abortive transcription rates.

Figure 46. Metagenes showing the Log2 ratio of genes after and before tetracycline induction from the second replicate with higher sequencing depth. (A) HBB with the Tet operon in the 5’UTR, (B) HBB with the Tet operon in intron 1, (C) HBB with the Tet operon inside exon 3 and (D) HBB with the Tet operon in the 3’UTR. For all samples, excluding C, this appears as expected, with enrichment of biotinylated Pol2 at the TSS after tetracycline induction. This appears similar to my results in Figure 28.
I measured this difference by calculating the \( \log_2 \) fold ratio of tetracycline induced compared to not induced at the promoters of protein coding genes on chromosomes 1-22. I then compared the average \( \log_2 \) fold ratios to see if they were significantly different between samples which had the Tet operon in different locations (Figure 47). Promoters were defined as the transcriptional start site +/- 250 base pairs. As previously mentioned, the sample with Tet operon from exon 3 of HBB transfected was excluded from the analysis. The remaining samples contained the Tet operon in the 5’ UTR, intron 1, and the 3’ UTR.

![Figure 47](image)

**Figure 47.** A boxplot of the \( \log_2 \) ratio of protein coding promoters (TSS+/− 250 bp) on chromosomes 1-22, after and before tetracycline induction for the second replicate. There is significant effect of Tet operon location on the \( \log_2 \) (ratio at the promoter) \( (p < 2.2e-16, \text{ one-way Anova}) \). Post-Hoc tests showed that each location was significantly different from each other, with a higher \( \log_2 \) (ratio at the promoter) progressing through the gene \( (P = 9.2e-9 \text{ comparing Tet operon at the 5’ UTR to intron 1, } P < 2.2e-16 \text{ comparing intron 1 to the 3’ UTR, all two-tailed } t \text{ tests}) \).

I found a significant effect of Tet operon location on the \( \log_2 \) (ratio at the promoter) \( (p < 2.2e-16, \text{ one-way Anova}) \). Post-Hoc tests showed that each location was significantly different from each other, with a higher \( \log_2 \) (ratio at the promoter) progressing through the gene \( (P = 9.2e-9 \text{ comparing Tet operon at the 5’ UTR to intron 1 and } P < 2.2e-16 \text{ comparing intron 1 to the 3’ UTR, all two-tailed } t \text{-tests}) \).
suggests that abortive transcription rate is highest at the 5’ end of a gene and decreases through the gene body.

One possible criticism of this method to quantify relative abortive transcription rates is that it assumes the pulldown efficiency between samples to be the same. For example, if the Tet operon in the 5’ UTR after tetracycline induction had a worse pulldown efficiency than all other samples then the same result could be observed. Whilst samples were processed in parallel, and reagents from the same batch were used to minimize the chance of this happening, it cannot be ruled out.

Therefore, I decided to reanalyse the first experiment as any difference in pulldown efficiency is likely to be technical. Whilst this experiment had worse signal to noise ratio, likely due to the much lower read depth, it might be possible to see if the same effect is present. If so, this provides stronger evidence for abortive transcription rates being higher towards the 5’ end of a gene. I followed the same protocol as before, I calculated the log2 fold ratio between tetracycline induced and not induced, at protein coding promoters on chromosomes 1-22 for each sample (Figure 48).
Again, I found that Tet operon location significantly affects the log2 fold ratio ($P = 2.5e^{-11}$, one-way Anova). Post-Hoc testing showed that Tet operon at the 5’UTR had the lowest log2 fold ratio (compared to the next lowest exon 3, $P = 0.00047$, two-tailed t-test) there was no significant difference between Tet operon in intron 1 and exon 3 ($P = 0.11$, two-tailed t-test), and the Tet operon in the 3’ UTR had the highest log2 fold ratio, which was significantly higher than the Tet operon at the 5’ UTR ($P = 2.2e^{-11}$, two-tailed t-test), Tet operon at exon 3 ($P = 0.00076$, two-tailed t-test) but not significantly different to Tet operon at intron 1 ($P=0.058$, two-tailed t-test).

Even with the caveats of a much noisier dataset, I found the same results. Therefore, I can conclude that the results from the larger dataset are unlikely to be due to a technical bias. Through inserting the Tet operon into different locations on a model gene, I found that there is greater abortive transcription in the 5’ UTR, lower in the gene body, and lower still in the 3’ UTR.

Figure 48. A boxplot of the log$_2$ ratio of protein coding promoters (TSS+/- 250 bp) on chromosomes 1-22, after and before tetracycline induction for the first replicate. Whilst differences between Tet operon locations are harder to see, Tet operon location significantly affects the log2 fold ratio ($P=2.5e^{-11}$, one-way Anova).
It would be interesting to investigate this effect of abortive transcription rates decreasing through a gene further, for example to test if it is a function of the gene length or gene region. An interesting experiment would be to analyse if an intron towards the end of a very long gene (e.g. DMD) has a lower abortive transcription rate than an intron in HBB, or an early intron in the same gene. Another question to answer is whether the 3’ UTR has a lower abortive transcription rate than the gene body, even if the 3’ UTR is on a shorter gene.

Mechanistically, it is interesting to speculate why this abortive transcription is reduced throughout the gene. A large number of factors are involved in productive elongation, including but not limited to pausing factors (e.g. NELF), processivity factors (e.g. Elongin), and kinases that phosphorylate key serine residues on the Pol2 CTD (e.g. Ser2-P, by P-TEFb), as well as various chromatin remodelling proteins (Kwak and Lis, 2013). Perhaps, as Pol2 progresses through a gene, stabilising factors accumulate along the transcription bubble. One study provides evidence for this (Fitz et al., 2018). Here, the authors depleted a pausing and elongation factor Spt5 and found that this only increased abortive transcription far from the promoter (15-20 kb), and therefore this protein only seems to act to reduce abortive transcription for long genes. This supports the evolutionary argument that increased pressures occur on reducing abortive transcription rates on longer genes as a protein exists which fulfils such a function.

Summary

Overall, I have used the assay that I developed in earlier chapters to measure abortive transcription along a model gene. I used Gibson assembly to synthesize the model gene HBB in an expression plasmid which contained the Tet operon inserted into different locations along the gene. I then transfected these plasmids along with plasmids encoding the fusion proteins Avi-RPB1 and TetR-BirA into Hela cells and used these to measure abortive transcription. I found significantly decreased abortive transcription as the model gene is progressed through, with the lowest abortive transcription rates in the 3’UTR and the highest at the 5’UTR, with the gene
body (intron 1 and exon 3) coming between. This evidence matches the evolutionary argument, that there is increased pressure to not abort transcription further through a gene due to more wasted resources producing a non-functional transcript. The evolutionary argument is also supported by separate research, showing that some elongation factors only reduce abortive transcription a long way into a gene (Fitz et al., 2018).
Conclusions

In this thesis, I have made 3 major discoveries. Firstly, I (in collaboration with my research group) discovered that 3′-5′ crosstalk contributes towards transcriptional noise (Cavallaro et al., 2021). I then developed a novel technology, which labels Pol2 by biotinylation at a specific genetic location, releases it with tetracycline induction and tracks the labelled Pol2 with a streptavidin pulldown. This technology was used to discover a mechanism behind 3′-5′ crosstalk in a model gene, which was found to be polymerase recycling, matching a microscopic model of transcription found in Cavallaro et al. (2021). I then used this technology I developed to measure a separate phenomenon, abortive transcription. I found that abortive transcription decreased as Pol2 progressed through a gene, with the highest rate of abortive transcription at the 5′ end, consistent with previous reports (Steurer et al., 2018), and abortive transcription reducing from the gene body to the 3′ end. I will discuss each of these findings and their implications in turn.

The finding that 3′-5′ crosstalk contributes to transcriptional noise has future implications. We found that a single base mutation on two model genes in the polyadenylation signal A₂TA₃ -> A₂GA₃, which disrupts this crosstalk, decreases transcriptional noise - for a given mean expression. This knowledge could be used to improve genetic circuits. For example, mutating the poly-adenylation signal could be used to ensure that the output of a cellular circuit for every cell is the same, with the same input. This could be useful in the field of biosensors (Bhalla et al., 2016) where it is important that the same input, will consistently give the same output.

Investigation of influenza infected cells (Russell et al., 2018) demonstrated that this decreased noise from the poly-adenylation site point mutation was not due to defective termination, as inhibiting termination upon influenza infection did not reduce the transcriptional noise. Furthermore, analysis of stranded RNA-Seq data (Cavallaro et al., 2021) showed that the poly-adenylation signal point mutation does not increase antisense transcription, which has previously been associated with buffering noise (Huber et al., 2016).
The effect of 3’-5’ crosstalk on transcriptional noise was measured on a genome wide scale. ChIA-PET datasets with a Pol2 pulldown (Heidari et al., 2014; Li et al., 2012) were analysed to calculate transcriptionally relevant 3’-5’ crosstalk. Transcriptional noise was calculated using a scRNA-Seq dataset (Klein et al., 2015) which was compared to 3’-5’ crosstalk. I found that genes with higher crosstalk had on average higher RNA expression and higher transcriptional noise, matching the results from the model genes study. Analysis of ChIP-Seq and ChIP-Exo experiments (McHaourab et al., 2018) further supported these results.

I investigated the function of genes with very high 3’-5’ crosstalk (top 10% of 3’-5’ interaction score – called looped genes). Across 3 repeats of K562 cell lines, and one MCF7 cell line, all with a Pol2 pulldown (Heidari et al., 2014; Li et al., 2012), I consistently found looped genes had the molecular function ‘structural component of the ribosome’ as the most enriched class. These genes encoded ribosomal protein genes, and they had significantly higher mean expression and higher transcriptional noise than other genes.

High mean expression was expected on account of their basal cellular function, but I had expected housekeeping genes like this to behave in a similar way across individual cells. I speculated that these genes need to be expressed at a high level which is why they are looped. Imaging data of another highly expressed gene class, those encoding ribosomal RNA (rRNA) which consist of more than 80% of total RNA may support this idea, as these genes have also been shown to be looped (Maiser et al., 2020). I also speculated that high transcriptional noise could be a disadvantageous trade-off for housekeeping genes. I found that looped genes in the class ‘structural component of the ribosome’ do have lower transcriptional noise than other looped genes, which suggests the cell uses other mechanisms to reduce the disadvantageous trade off. I also suggested that discovering looped ribosomal protein encoding genes may be a result of cell lines undergoing continual selection for higher growth rates, which require the fast production of ribosomes. This might not necessarily be true in vivo and would be interesting to test.
It would be interesting to see if genes have high 3'-5' crosstalk in a living system where transcriptional noise itself is selected for. An example of this could be a bet hedging strategy, where single cell organisms can employ a strategy of expressing different phenotypes in the same environment – which can give survivability in unpredictable environments, for instance bacterial persistence to antibiotics (Balaban et al., 2004). Noisy gene expression can be used to generate a bet-hedging strategy. For instance, the alternative stress response sigma factor RpoS is expressed in a noisy manner across *E. coli* and can help survivability to peroxide treatment (Patange et al., 2018). In a similar manner, in the eukaryotic *S. cerevisiae* a stress protectant gene Tsl1 also has high variability, due to a combination of stochastic and deterministic factors, which help improve survivability to heat killing (Levy et al., 2012). Indeed, like eukaryotes, prokaryotes also have a complex three-dimensional genome containing long range contacts (Lioy et al., 2018). Therefore, measurements of 3'-5' crosstalk in systems such as these, combined with scRNA-Seq to measure transcriptional noise would be interesting to see if genes with high crosstalk and noise are enriched among stress responsive genes.

I then investigated the interactions of the three-dimensional genome, treating these interactions as a network. I treated each 2 kb region from the Hi-C style contact matrix of the genome as a node and I defined interactions between nodes as links. I found that most nodes interacted with few others, but some had many interactions – much more than predicted by a random network structure. Plotting the cumulative fraction of nodes against the degree (number of connections to other nodes) on a logarithmic scale showed that this network appeared scale-free. A key feature of scale-free networks is their resistance to random attacks, which may be advantageous for the three-dimensional genome. I found that hubs, defined as nodes with a high degree were enriched in both genes and enhancers. This provides a further evolutionary explanation of the three-dimensional genome structure. For example, enhancers being found in hubs can explain how an enhancer can regulate multiple distal genes, and genes being found in hubs can point towards multiple elements contributing to their regulation.
I developed and tested a system where the large subunit of RNA polymerase 2 is fused to the avitag (Avi-RPB1) and biotinylated by a fusion protein of the Tet repressor and a biotin ligase (TetR-BirA). I demonstrated that this biotinylation occurs specifically at the Tet operon and requires all three components of the system to be present to work (Avi-RPB1, TetR-BirA, Tet operon). I found, through ChIP-sequencing using a biotin pulldown, that biotinylated Pol2 behaved similarly to normal Pol2, as they share similar localisations. I originally hoped that this system could be used to directly measure polymerase recycling by tracking a Biotin-Pol2 signal, maintained on HBB after tetracycline induction, and that the differences in maintenance times between the WT and pA- cells could then demonstrate polymerase recycling.

Unfortunately, the system did not work entirely as I suspected, as Biotin-Pol2 appeared across the genome even before tetracycline induction. This meant that the presence of Biotin-Pol2 at HBB after tetracycline induction could either be due to recycling or continued initiation, meaning this approach alone cannot be used to measure recycling.

I then developed a simplified mathematical model of my system which demonstrated that the ratio of biotinylated Pol2 to total Pol2 at HBB compared to a control gene (GAPDH) can be used to measure if polymerase recycling or initiation are reduced when the poly-adenylation signal is mutated in the HBB gene. ELISA and mass spectrometry experiments showed that both cell lines had the same starting conditions in the total concentration of Pol2 and Biotin-Pol2.

A ChIP-qPCR experiment showed that after induction this ratio is higher in the WT line compared to the pA- line. This demonstrated that the mechanism behind 3’-5’ crosstalk contributing to transcriptional noise is polymerase recycling, which is disrupted with the poly-adenylation signal mutation. Of course, there is still some open questions here, such as how a burst of transcription ends. Suggestions of this include ideas that this 3’-5’ crosstalk is transient in nature or other transcription
factors dissociating from a gene (Hebenstreit, 2013). Both suggestions would result in disruption of polymerase recycling and stopping the burst of transcription.

Knowledge of the mechanism that polymerase recycling is behind 3’-5’ crosstalk contributing to transcriptional noise can also be a fruitful avenue for future research. For example, it might be possible to synthesise a system to increase or decrease recycling, potentially by forcing a gene to loop in such a way to increase or decrease the three-dimensional distance between the TSS and TES. A common feature of DNA loops is that they are marked by both CTCF in a convergent orientation and Cohesin (Rao et al., 2014). Whilst only a small fraction of possible CTCF binding pairs interact to form a loop (2-5%) (Rao et al., 2014; Tang et al., 2015; Xi and Beer, 2021), looping interactions can be predicted by a relatively simple loop extrusion model combined with competition between overlapping loops (Xi and Beer, 2021). This could mean that experiments can be done in-silico to predict useful genomic sites to insert CTCF binding pairs to bring the TSS and TES closer together or further apart to facilitate or disrupt polymerase recycling. This could in turn alter the noise profile, and expression of genes. For example, in a system where it is important to express a gene to maximal capacity, perhaps to manufacture a useful protein, total yield is important, which could be increased by directly creating a loop, forcing an increase in polymerase recycling. Furthermore, in systems where increased transcriptional noise is wanted, for example stochastic computation (Grozinger et al., 2019), then adding a loop may also be beneficial.

I then used the system that I developed to measure a separate phenomenon, that of abortive transcription. I measured relative abortive transcription at a model gene (HBB), where I synthesised different versions containing the Tet operon at different genic locations across this gene. I reasoned that with higher abortive transcription, the difference in number of reads at other genes’ promoters between tetracycline induced and not induced would be reduced.

I found that the rate of abortive transcription, measured in this way appears to reduce through a gene, with the rates of abortive transcription highest at the 5’ end,
and lowest at the 3’ end. This matches an evolutionary argument of energy wastage producing potentially non-functional partial mRNAs and may help explain how long genes such as DMD are transcribed. Mechanisms may involve proteins similar to Spt5, which was found to reduce abortive transcription, but only after around 15-20 kb from the promoter.

This method can be improved further, to reduce release from the Tet operon before tetracycline induction. For example, it may be possible to minimise abortive transcription, potentially through inserting the Tet operon towards the end of a long gene and combining this with reducing the time TetR-BirA is expressed to reduce the opportunity for abortive transcription. This could be combined with a stable integration of Avi-RPB1 to ensure every cell is expressing this fusion protein, possibly decreasing the fraction of other non-biotinylated proteins which are pulled down with streptavidin.

A hypomorphic mutation of BirA (G115S) could also be used to reduce biotinylation rates (Schlissel and Rine, 2019), which may reduce off target biotinylation or biotinylation of more transient polymerases (i.e. polymerases which may be near the Tet operon, but not bound to DNA and initiating transcription). A second approach could be engineering two separate TetR fusion proteins, in which BirA is split into two (Schlissel and Rine, 2019). This means that BirA could only biotinylate Pol2 if the two different versions of TetR-BirA form a heterodimer, again likely to reduce off-target biotinylation. If, while unlikely, subsequent experiments (Part 2, Chapter 4) show leaky expression through the Tet operon, the leaky expression could be reduced by adding an array of multiple Tet operons, rather than the two in my system.

With improvements, resulting in minimal release from the Tet operon before tetracycline induction, the system I have developed could be adapted to multiple other systems. One possible example could be the study of three-dimensional genome architecture. Here, when biotinylated Pol2 is released from the Tet operon by tetracycline it spreads across the genome and is likely to arrive at nearby locations first – matching my tentative conclusions from Part 2, Chapter 1. Using multiple
release points and timing arrival times at different locations could be used for triangulation of genome topology.

Other applications of this technique can be envisioned; as the system is modular, it is possible to simply change the fusion protein containing the Avitag to investigate a different factor. For example, different polymerases (I and III) could be tracked to investigate non mRNA transcription. Furthermore, BirA could be switched for a promiscuous biotin ligase, to discover all proteins interacting at the specified genetic location, in a similar fashion to BioID (Roux et al., 2012). This could even be extended to using dCas9 fused to a promiscuous biotin ligase, allowing biotin labelling of all proteins which localise to a specified genetic location. It is also theoretically possible to adapt this system to the single cell level in a similar fashion to CUT&TAG (Kaya-Okur et al., 2020) to explore cell to cell variability in polymerase dynamics.

I hope the community will find the techniques I developed and my findings interesting, and I look forward to seeing the next steps.
Works cited


Dublanche, Y., Michalodimitrakis, K., Kümmerer, N., Foglierini, M., and Serrano, L.


Fitz, J., Neumann, T., and Pavri, R. (2018). Regulation of RNA polymerase II processivity by Spt5 is restricted to a narrow window during elongation. EMBO J.


Hsiao, L.L., Dangond, F., Yoshida, T., Hong, R., Jensen, R. V., Misra, J., Dillon, W., Lee,


Mueller, F., Senecal, A., Tantale, K., Marie-Nelly, H., Ly, N., Collin, O., Basyuk, E.,


R Core Team (2020). R: A Language and Environment for Statistical Computing.


Tiberi, S., Walsh, M., Cavallaro, M., Hebenstreit, D., and Rbel FinkenstäDt, B. Bayesian inference on stochastic gene transcription from flow cytometry data.


