Insights into transcription through nascent RNA sequencing

by

Artur Botelho Veloso

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2014

Doctoral Committee:

Professor Mats E.D. Ljungman, Chair
Associate Professor Scott E. Barolo
Professor Daniel M. Burns Jr.
Assistant Professor Hui Jiang
Professor Kerby A. Shedden
Associate Professor Thomas E. Wilson

Para os meus pais, Léa e Rogério.

ACKNOWLEDGEMENTS

Até onde eu me lembro, o meu interesse em ciência e pesquisa me acompanhou durante toda a minha vida. Quando criança eu me sentia fascinado pelo mundo científico e almejava o dia em que eu faria parte desse mundo. Durante a última década eu tive a oportunidade de perseguir esse sonho, mas isso foi feita a duras penas. A minha decisão de mudar de país e deixar família e amigos para trás não foi tomada facilmente. Eu não a teria tomado, no entanto, se não tivesse recebido tanto apoio dessas mesmas pessoas. Esse processo não foi fácil para mim, assim como não foi fácil para eles. Por isso, eu primeiramente gostaria de agradecer às pessoas que tem me apoiado durante toda minha vida. Minha mãe e meu pai, Léa e Rogério, foram fundamentais em meu desenvolvimento. Além das qualidades básicas que bons pais ensinam a filhos, eles me ajudaram a desenvolver um forte sentido de independência. Essa independência foi fundamental na minha decisão de estudar em outro país. Outro aspecto importante nessa decisão foi a minha ambição. Uma das pessoas que tiveram o maior impacto nessa característica foi meu irmão, Cristiano. Vários outros familiares me apoiaram durante minha vida, como minhas avós Ana e Maria, e vários tios e tias, especialmente meu tio Milton. Apesar de terem ingressado em minha vida há menos tempo, minha cunhada Jennifer e minha sobrinha Sofia já afetaram fortemente minha vida. A minha formação pessoal tabém foi fortemente influenciada pelos meus amigos Marcelo Moura, Mateus Dutra, Felipe Reis, Henrique Amaral e Leonardo Amaral. A todos vocês, obrigado pelo apoio e carinho.

On April 22nd, 2011, I had a meeting with Dr. Mats Ljungman to discuss the possibility of doing a research rotation in his laboratory. In that meeting he introduced me to the concept of nascent RNA sequencing, and showed the brand new data that they had generated using Bru-seq, BruChase-seq and BruUV-seq. During the three years that followed I've had the opportunity to develop analysis techniques for these and other projects. Much of the computational work necessary for the development of such techniques was carried out under the supervision of Dr. Thomas E. Wilson. Both Mats and Tom have been essential in my progress these last years and I'm very thankful to have had their support. While in the Ljungman lab, I had the opportunity to work with very talented people. First and foremost, Michelle Paulsen was not just amazingly efficient at generating close to all the data I used in this thesis, she was also a great friend. Other students and researches in the lab also greatly helped me during this process. Brian Magnuson, Leonardo Lima and Killeen Kirkconnell were extremely helpful in helping me broaden my understanding of molecular biology and transcription. It was also very enjoyable to initiate into the field of bioinformatics other students in the lab such as Nathan Berg, Hailey Lefkofsky and Karan Bedi.

Prior to joining the University of Michigan, I had no formal training in quantitative sciences. In spite of that, the Program in the Biomedical Sciences and the Bioinformatics Graduate Program accepted me as a student and gave me free rein to experiment with different classes and laboratory rotations. I'm grateful to Dr. Margit Burmeister, Dr. Dan Burns and Dr. Kerby Shedden for their help and advice in those initial years. I'm also thankful for the insightful discussions that happened during my committee meetings. During these, Mats Ljungman and Tom Wilson were joined by Dan Burns, Kerby Shedden, Dr. Hui Jiang, and Dr. Scott Barolo.

Finally, I'm grateful for all my friends and colleagues in Ann Arbor and elsewhere in the world who gave me support and helped me along the way. Within my cohort at the Bioinformatics Graduate Program I met extremely friendly and warm people who kept me company for these last years. Among that group, people such as Kraig Stevenson, Mallory Freeberg, Shanshan Cheng, Avinash Shanmugam, and Ellen Schmidt, came to be some of my dearest friends. It would take too long to describe how other bioinformatics classmates, colleagues, collaborators, and other people came into my life and the impact they caused. To name a few in alphabetical order: Thomas Baird, Alejandro Balbin, Bruna de Castro, Juliana Chevitarese, Jeremy Doody, Melissa Eslinger, Lucas Faissal, Jennifer Fountain, Elsie Grace, Joann Gruber, Jinyi Li, Sunit Jain, Andrew Kocab, Patrick Harrington, Yongsehgn Huang, Kathryn Iverson, Lindsey MacDonald, Steven O'Connell, Felipe Rozenberg, Paul Tamoshunas.

To all of the people named above, thank you!

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF ABBREVIATIONS

4-thiouridine	4sU
BrdU	Bromodeoxyuridine
Bru	bromouridine
ChIP	Chromatin Immunoprecipitation
CPE	core promoter element
CPT	camptothecin
CTD	C-Terminal domain
DNA	Deoxyribonucleic acid
DRB	5,6-dichlorobenzimidazole 1-β-D-ribofuranoside
DSIF	DRB sensitivity inducing factor
eRNA	enhancer RNA
GRO-Seq	Global run-on sequencing
HMM	Hidden Markov Model
mRNA	Messenger RNA
NELF	Negative elongation factor
NET-seq	Native elongating transcript sequencing
PIC	pre-initiation complex
PROMPT	Promoter upstream transcript
RNA	Ribonucleic acid
RNAPII	RNA polymerase II
RPKM	reads per kilobase per million mapped reads
TF	transcription factor
TNF	Tumor Necrosis Factor
Top1	DNA topoisomerase I
TSS	transcription start site
UV	Ultraviolet

CHAPTER I

Introduction

1.1 Research overview

One of the most important fields in modern molecular biology is the study of RNA, or transcriptomics. Historically, the central dogma of molecular biology has been summarized as "DNA makes RNA and RNA makes protein". Therefore, by quantifying RNA production, one is can also infer protein abundance in the cell. While that's a very important aspect of transcriptomics, the RNA world is much greater than that. Research in the past decades has demonstrated that there are several different RNA populations, which have very specific roles inside cells.

Most researches, however, are interested mostly in protein coding genes. Because of that, they focus on polyadenylated RNA. Conventional transcriptome studies (e.g. ESTs, gene expression microarrays, RNA-seq) have been successfully used from gene identification and annotation to finding disease biomarkers. Since these studies focus on the mature form of transcripts, they tend to overlook much of the complexity involved in gene regulation. Regular transcriptome studies quantify the abundance of transcripts at a given time point. They do not, however, determine how those values are reached. There are two non-exclusive ways that transcript abundance can be reached:

Synthesis One possible way to regulate the abundance of a transcript is to determine how much of that RNA is transcribed in the system. Assuming that all stays the same, a greater synthesis of a given transcript will eventually cause its abundance to be increased.

Stability The degradation of molecules in a biological system is a natural occurrence. Not all transcripts, however, are degraded at the same rate. If two transcripts are synthesized at the same rate but degraded at different rates, their abundance will differ.

In order to better understand gene regulation, it is necessary to be able to discern between transcript synthesis and transcript stability. This, however, is impossible to do with conventional transcriptomics studies. The first step to accomplish this goal is to be able to measure the synthesis of transcripts. This can be achieved by measuring the rate at which RNA is transcribed during a short period of time. The newly made RNA is referred to as nascent RNA. Two methods could potentially be used to determine the stability of transcripts. The most straightforward method would be to measure a transcript's abundance in the nascent RNA and in the total pool of RNA. RNA stability could be calculated as a function of abundance of a transcript in the nascent and in the total pools of RNA. For example, an unstable transcript could display high abundance in the nascent RNA and low abundance in the total RNA pool. If homeostasis was disrupted, however, this technique would falter. Since the stability measurements are based on the total pool of RNA, changes in transcript stability would only be measurable once enough time had passed for the total pool of RNA to be affected. Therefore, short-term stimulus-induced changes in stability would be difficult to measure.

This problem can be circumvented if, instead of studying the complete RNA pool,

one is able to focus on a limited pool of RNA that was transcribed synchronously. This leads us to the second method to determine transcript stability. The abundance of a transcript in the nascent RNA can be compared to its abundance in the same RNA population after a given amount of time has passed. This allows for stability measurements to be carried out in homeostasis or after a treatment.

The effect of treatments on gene expression can also be more quickly observed when analyzing changes to the nascent RNA. When using standard techniques, such as RNA-seq with polyadenylated transcripts, treatment-induced changes in the synthesis of transcripts can only be observed once they have accumulated enough to change the total pool of RNA. This is particularly relevant for large genes which might take a long time to be transcribed. The treatment might induce immediate change in the synthesis of the transcript, but that will go completely unnoticed until the whole transcript is transcribed and polyadenylation takes place. Therefore, nascent RNA sequencing allows for changes in transcript expression to be measured much sooner.

This is particularly important when the focus of interest is not the final changes in transcript abundance, but the effect of the treatment on the transcription process. There are several drugs (e.g. camptothecin, etoposide) and environmental factors (e.g. UV radiation, reactive oxygen species) that directly or indirectly affect transcription. By analyzing the nascent RNA, it is possible to observe and better determine how the treatment is changing transcription.

These treatments can also be used to better understand details about transcription. Transcription is a complex, multistage process that involves a very large number of proteins. Nascent RNA sequencing can allow us to observe how transcription is affected in the absence of any one of these proteins. Also, more complex experimental

designs can be used to answer questions about the individual steps of transcription. For example, from how many different promoters does transcription initiate? What is the elongation speed of RNA polymerase? Does transcription terminate always at the same place?

In this thesis I used data generated from a set of techniques that have been developed by Dr. Mats Ljungman's research group as a tool to study nascent RNA expression and the transcription process as a whole. I collaborated with Dr. Thomas E. Wilson in establishing a pipeline for mapping and carrying out initial analysis of this data. Several of the research questions addressed in this thesis could not be solved with existent software. Therefore, most of my efforts involved developing computational and statistical analysis approaches that could answer those questions.

1.2 Dissertation outline

This dissertation is divided into seven chapters. Chapter I provides an introduction to the major themes discussed throughout the thesis, while chapter VII summarizes the major findings, discusses their relevance and how this research could be expanded upon. Chapters II to VI are the scientific reports that were generated based upon the research carried out by myself and other laboratory members.

Our research group developed a technique based on metabolic labeling of RNA using bromouridine (Bru) in order to study nascent RNA synthesis. This technique was named Bru-seq. One of the main advantages of this approach is that it allowed for a chase to be carried out after the labeling, which made studying RNA stability and RNA post-transcription processing possible. The name given to this technique was BruChase-seq. The experimental procedures necessary to carry out these techniques and a thorough description of the analysis pipeline used with this data are described

in chapter II. The same chapter also gives examples of how transcript synthesis, stability and processing can be measured under homeostasis using these techniques.

Biological systems tend to change and adapt once homeostasis is disturbed. It is well known that treatment-induced changes in gene expression can occur, but it is not clear what proportion of these changes occur as a result of changes in the rates of synthesis or stability. In chapter III, different aspects of transcription regulation are tested by analyzing the effect of the proinflammatory cytokine Tumor Necrosis Factor (TNF) in transcript synthesis and stability in human fibroblasts. The results, show coordinated regulation at both the level of synthesis and stability.

The development of a technique which allows the study of nascent RNA makes it possible to explore details of the RNA synthesis process. Several different approaches could be taken, but we decided to focus on treatments that inhibit transcription at one of its stages. In chapters IV to VI we explore three of these treatments and use them to increase out knowledge of transcription.

UV radiation causes DNA lesions, which work as a barrier for elongating DNA and RNA polymerases. In chapter IV we use the Bru-seq technique in cells that were irradiated to understand exactly how transcription is affected. We observed that the transcription signals were redistributed from bodies of genes to TSS and putative enhancers, which is reasonable since elongation is inhibited but initiation is unaffected. Due to the usefulness of accumulating signal around TSS and enhancers, this chapter describes how this treatment could be used as a technique, referred to as BruUV-seq.

Another treatment that inhibits transcription elongation is the antitumoral drug camptothecin (CPT). While its mechanism of action is completely different from UV radiation, we found that the effect it has on nascent RNA is fairly similar. In chapter

V we explore the effect of CPT on transcript synthesis during and 15 and 30 minutes after treatment. Due to CPT's inhibition of elongation and the inability of blocked RNAPII to resume transcription after CPT removal, transcript synthesis recovery is delayed in larger genes. This might partially explain the antitumoral activity of CPT since it leads to the inhibition of large proto-oncogenes and large anti-apoptotic genes.

Transcript synthesis recovery after CPT demonstrated how important gene length can be to the expression of a gene. Similarly, the speed of elongation of RNA polymerase is also very important to determine the time taken for transcription to be completed. In chapter VI we describe a technique for measuring genome-wide transcript elongation rate called BruDRB-seq. While the gene's elongation rate can be quite different, we observed a correlation in the elongation rate of transcription across genes in five cell lines.

Since chapters II to VI were prepared for publication in scientific journals, they do not contain detailed background information. The next sections of this chapter contain information that will aid the understanding of those chapters.

1.3 RNA transcription

RNA transcription is a very complex process, which is usually divided into three steps, transcription initiation, transcription elongation and transcription termination. Initially, RNA polymerase II (RNAPII) and a wide range of general transcription factors form a functional pre-initiation complex (PIC) and bind to the gene's promoter (see 1.3.1). Once bound, RNAPII transcribes a small number of nucleotides (5-25). This stage, transcription initiation, may lead to productive elongation but could lead to abortive transcription (see 1.3.2). If transcription is not aborted and

RNAPII is released from the promoter it can move along while transcribing the DNA strand, which is referred to as transcription elongation (see 1.3.4). Most of the transcript processing, such as capping and splicing, happen while elongation is taking place. The dissociation of RNAPII from the DNA molecule characterizes transcription termination (see 1.3.5).

1.3.1 Formation of pre-initiation complex

The eukaryotic DNA molecule is highly organized and protected by histones and other scaffolding proteins. Specialized proteins are necessary to aid in the identification of the functional elements that lie in the genome. This identification happens on the basis of conserved sequences that usually lie close to the promoter region of the gene (Juven-Gershon et al., 2008). These core promoter elements (CPE) may contain many different motifs (e.g. BRE, Inr, MTE, DCE). The TATA box is one of the most studied examples of CPE. It is also one of the most ancient conserved DNA sequences, present from Archaea through eukaryotes (Reeve, 2003).

The TATA box is recognized and bound to by TATA-box-binding-protein. Binding to the TATA box leads to a change in conformation of the DNA molecule, which makes it possible for other transcription factors (TF), such as TFIID, to bind to the promoter region (Shandilya and Roberts, 2012). In the absence of the TATA box, TFIID and other transcription factors will bind to other CPE, such as Initiator and Downstream Promoter Element (Baumann et al., 2010). Several CPE are flanked by TFIIB recognition elements, which are binding sites for transcription factor TFIIB. The binding of TATA-box-binding-protein to the DNA molecule is stabilized by the binding of TFIIB to the complex. The PIC is further stabilized by binding to RNAPII and TFIIF (Deng and Roberts, 2007).

1.3.2 Transcription initiation

Formation of a stable PIC does not guarantee that transcription initiation will take place. In some cases, transcription is aborted after RNAPII produces a transcript that is approximately 5 nucleotides long (Saunders et al., 2006). Abortive initiation is related to structural changes that occur in the RNAPII-DNA complex when the first nucleotides are polymerized. It has been suggested that abortive initiation could act as a cellular checkpoint to avoid nonspecific transcription (Liu et al., 2011). As more nucleotides are added to the polymerizing RNA molecule, the tendency for a successful promoter escape increases. At approximately 10 nucleotides, the likelihood of abortive initiation happening greatly decreases (Holstege et al., 1997). At this time, the ATP and TFIIH requirements for the reaction end and the transcription bubble collapses (Saunders et al., 2006). A reduction in upstream transcript slippage, the pairing of the RNA molecule with an upstream DNA sequence, also occurs approximately after the polymerization of 10 nucleotides (Pal and Luse, 2003). Once the RNA molecule reaches 25 nucleotides it is considered that the transition from transcription initiation into transcription elongation was successful. Roughly at that length, the cap structure is added to the 5' end of the transcript (Rasmussen and Lis, 1993).

In order for the capping to take place it is necessary for the fifth Serine of the RNAPII C-Terminal domain (CTD) repeats to be phosphorylated (Komarnitsky et al., 2000). Post-transcriptional modifications to the CTD are extremely important for a successful transcription initiation. In humans, the CTD is composed of 52 repeats of the peptides Tyrosine (Y) - Serine (S) - Proline (P) - Threonine (T) - S - P - S. The most common modifications that the CTD can be subjected to are: peptidylprolyl isomerization of the Proline amino acids; glycosylation of Serine

and Threonine; phosphorylation of Serine (Egloff and Murphy, 2008). The Serine residues in the CTD of a RNAPII are usually hypo-phosphorylated when the enzyme is recruited to form a PIC. A very important stage of transcription initiation is the phosphorylation of Serine 5 (Chapman et al., 2007). Post-transcriptional modifications of other proteins that make up the PIC also seem to be essential for a successful transcription initiation. For example, in certain genes it is essential for Ser65 of protein TFIIB to be phosphorylated (Wang et al., 2010).

1.3.3 Promoter-proximal pausing

A successful transcription initiation does not necessarily guarantee an immediate transition into transcription elongation. In approximately 30% of human genes, RNAPII pauses after transcribing between 20 and 60 nucleotides. This pausing is transient, which allows the RNAPII to resume transcription elongation (Adelman and Lis, 2012). Evidence for promoter-proximal pausing was initially observed using UV protein-DNA crosslinking (Gilmour and Lis, 1986). The importance of this event was only appreciated when it was observed to take place in a large number of genes using genome-wide techniques such as RNAPII Chromatin Immunoprecipitation (ChIP-chip) (Kim et al., 2005) and Global nuclear run-on sequencing (GRO-seq) (Core et al., 2008) assays. The processes of pausing and release are regulated by two factors. They are DRB sensitivity inducing factor (DSIF) (Wada et al., 1998a) and negative elongation factor (NELF) (Yamaguchi et al., 1999). Release from pausing is accomplished by phosphorylation of the DSIF/NELF complex by P-TEFb (Wada et al., 1998b).

While it is clear that promoter-proximal pausing occurs in a wide range of eukaryotes, its function is still not well understood. The authors Adelman and Lis (2012) propose four non-exclusive models for its function: (1) The presence of a

stalled RNAPII leads to a nucleosome depleted promoter, which improves the binding of transcription factors; (2) Certain gene expression programs depend on rapid response to stimuli. Changes in gene expression can be expedited by skipping the PIC recruitment and transcription initiation stages. Therefore, genes with paused RNAPII would be able to be activated more swiftly; (3) Promoter pausing is another regulatory step in gene expression; (4) Due to co-transcriptional processing, RNAPII is associated to several other proteins during transcription. Promoter pausing could function as a checkpoint to determine if all necessary protein complexes were coupled to RNAPII.

1.3.4 Transcription elongation

Transcription elongation starts when RNAPII is able to move away from the promoter and into the body of the genomic feature being transcribed. As indicated in 1.3.2, post-transcriptional modifications of RNAPII's CTD is very important for RNA transcription. The conventional view is that Serine 5 phosphorylation is high around the TSS and, as RNAPII progresses through the gene, Serine 5 phosphorylation is lost while Serine 2 phosphorylation is gained (Egloff and Murphy, 2008). Phosphorylation of Serine 7 seems to follow the same pattern as the observed in Serine 5 (Glover-Cutter et al., 2009). Similarly, to Serine 2, Threonine 4 also is phosphorylation towards the 3' end of genes (Hintermair et al., 2012).

A large portion of the transcription elongation literature seems to heavily focus on the release from promoter-proximal pausing (Peterlin and Price, 2006). While this transition is very important, such studies tend to overlook the complexity of the events that take place while RNAPII is engaged in elongation. For example, RNAPII does not move at a constant pace during transcription elongation. Not only does the elongation rate vary, but the enzyme can pause during elongation for several minutes

(Darzacq et al., 2007a). These changes in elongation rate can be very important for the fate of the RNA molecule. Co-transcriptional alternative splicing can be caused by changing the elongation rate of RNAPII (Shukla and Oberdoerffer, 2012). This can be achieved by causing mutations to the CTD in RNAPII that lead to slower elongation rates (de la Mata et al., 2003) or by adding a pausing sequence a gene (Roberts et al., 1998).

During elongation, RNAPII can pause at many sites within a gene. The pausing seems to happen most frequently at an adenine nucleotide, which is usually followed by a thymine and then a guanine (Churchman and Weissman, 2011). Pausing is usually associated with backtracking of RNAPII and cleavage of the RNA molecule. In order to backtracking take place, it is necessary to destabilize the 3-proximal RNADNA hybrid (Nudler et al., 1997). The RNA-DNA hybrid is between 8 and 9 base pairs long and is the most important molecule to maintain the stability of the RNAPII elongating complex (Kireeva et al., 2000). The length of the RNA-DNA hybrid is of extreme importance. One it is shortened, RNAPII changes from transcription elongation into transcription termination (Komissarova et al., 2002).

1.3.5 Transcription termination

Accumulation of Serine 2 phosphorylation on RNAPII's CTD is extremely important for a successful transcription termination. In the absence of such phosphorylation, enzymatic complexes necessary for proper processing of the 3' end of the RNA molecule are not recruited (Ahn et al., 2004). A very large number of proteins, more than 80, are either directly involved or associated with pre-mRNA 3' processing (Shi et al., 2009). Depending on the biological organism and/or the type of molecule being transcribed, there are several different pathways that are used to achieve transcription termination. In eukaryotes, termination can happen through a

Sen1-dependent or a Poly(A)-dependent pathway. Processing of small nucler RNA and small nucleolar RNA is carried out by the Sen1-dependent pathway and does not lead to polyadenylation. The Poly(A)-dependent pathway, on the other hand, is used in the processing of mRNA and leads to polyadenylated molecules (Kuehner et al., 2011).

1.4 Nascent RNA technologies

As discussed in 1.1, in order to carry out the work presented in this thesis, it was necessary to study the nascent RNA expression. This was accomplished by exposing cell cultures to bromouridine for a set amount of time. Bru is incorporated into RNA molecules that are being synthesized during the labeling period. These molecules can be isolated using anti-Bromodeoxyuridine (BrdU) antibodies. Analysis of these molecules will provide information regarding RNA transcription and any RNA processing that might have taken place during the labeling period. In addition, including a chase period in which the Bru is washed out and cells are incubated in a high concentration of regular uridine, allows us to study the fate of the nascent RNA over time. Depending on how long the pulse-chase lasts, analysis of the reminiscent labeled molecules can be used to learn more about different steps of RNA processing. For example, a short chase might be helpful to understand RNA splicing, while a longer can be used to measure RNA stability. The technique used in this thesis is thoroughly described in chapters II and III. In the next sections we will discuss other available techniques that could be used on similar studies.

1.4.1 GRO-seq

The most widely used nascent RNA sequencing technique is global run-on sequencing (GRO-seq) (Core et al., 2008). Nuclear run-on assays were used to mea-

sure overall or gene-specific transcription rates for a very long time prior to the development of GRO-seq (Hirayoshi and Lis, 1999). Nuclear run-on techniques are based on isolating the cell's nuclei and then allowing polymerases to move through the DNA molecule. This happens in the presence of the detergent sarkosyl, which inhibits the coupling to the DNA of new polymerases, but does not interfere with the bound molecules (Gariglio and Mousset, 1975). Transcription elongation is allowed to happen in vitro in the presence of a labeling agent, usually radioactive uridine or bromouridine. Since the reaction happens in vitro, there is very little RNA degradation during the labeling phase. Therefore, unstable RNA molecules, such as enhancer RNA (eRNA), can be detected in greater extent than in total RNA (Core et al., 2012). GRO-seq can also be used to gain insight into the transcription process by interfering with gene expression prior to labeling. RNAPII transcription elongation rates have been measured using GRO-seq by activating a cellular signaling pathway (Danko et al., 2013) or by exposing the cells to flavopiridol (Jonkers et al., 2014). Since the labeling happens in vitro, however, it is impossible to use GRO-seq to analyze post-transcription RNA processing such as splicing and stability.

1.4.2 NET-seq

A lot can be learned about RNA transcription based on the distribution of RNAP throughout the genome. RNAP ChIP-seq, for example, can be used to determine the distribution of the complex through chromosomes, but the data does not indicate which strand is bound by the complex and if there is active transcription happening (Lefrançois et al., 2009). In order to determine the position of actively elongating RNAP, native elongating transcript sequencing (NET-seq) was developed. Since the DNA-RNA-RNAP ternary complex is extremely stable (Cai and Luse, 1987), it is possible to immunoprecipitate the whole ternary complex using antibodies specific

for RNAP. Sequencing of the 3' ends of the isolated RNA molecules indicates the position of RNAP at a nucleotide resolution. This technique was used to demonstrate that there are a very high number of RNAP pause sites in yeast. Interestingly, a high level of conservation was observed in the sequence of the identified pause sites, indicating that these are not random events (Churchman and Weissman, 2011). Since NET-seq focuses on RNA that is still bound to the ternary complex, it enriches for unstable RNA molecules. Transcription can occur on the non-canonical direction of a promoter, leading to the production of PROMPT or cryptic unstable transcript. A non-essential gene deletion screening was used in combination with NET-seq to determine that cryptic unstable transcript production is facilitated by H3K56 hyperacetylation (Marquardt et al., 2014).

1.4.3 Nascent-seq

Similarly to NET-seq, Nascent-seq is also based on the high stability of the DNA-RNA-RNAP ternary complex. For Nascent-seq, however, this stability is explored differently. The nuclei of cells are extracted and fractionated into a pellet and a supernatant. The pellet contains the DNA, histones and ternary complexes, while the supernatant contains the nonhistone proteins and RNA molecules that are not attached to a ternary complex (Wuarin and Schibler, 1994). Since the RNA present in the pellet was transcribed in vivo and is still associated to RNAPII, any modifications to it must have occurred cotranscriptionally. Using Nascent-seq, it was estimated that up to 13% of introns in *Drosophila* have poor cotranscriptional splicing. Interestingly, cotranscriptional splicing was more efficient in cells expressing a RNAPII mutant which has lower elongation rate (Khodor et al., 2011). Another type of RNA processing is the chemical modification of adenosine into inosine by the RNA editing enzyme ADAR (Kim et al., 1994). Nascent-seq data indicated a correlation in the

amount of modification observed in nascent and mature RNA, demonstrating the cotranscriptional nature of this processing event. Even though these modifications are usually enriched in exonic regions in comparison to intronic regions, the same study found that introns that were poorly spliced had a disproportionate amount of adenosine into inosine modifications (Rodriguez et al., 2012).

1.4.4 Metabolic labeling

The research described in chapters II to VI is based on metabolic labeling of nascent RNA with Bru (Paulsen et al., 2013b,a). Since this approach does not depend on complex protocols that involved enucleation of cells (such as GRO-seq and Nascent-seq) or immunoprecipitation of the ternary complex (such as NET-seq), it has been implemented by several different research groups. The basic idea is to use bromouridine or 4-thiouridine (4sU) to label the RNA produced during a given time frame. Such approach was used to compare RNA populations between asynchronous and G2-arrested cells. Differential expression was observed both at the nascent and the mature RNA level, but there was only a small overlap in genes differently regulated across these two RNA pools (Ohtsu et al., 2008).

Transcript decay can be computated by comparing the abundance of a transcript in the nascent and in the mature RNA pools. The response to lipopolysaccharide by bone marrowderived dendritic cells was assessed both at the nascent and mature levels. While transcript synthesis was the main factor determining transcript abundance in the mature RNA pool, transcript degradation was very important to enable sharp changes in the mature RNA levels (Rabani et al., 2011). As it would be expected, impairment of RNA transcription and mRNA degradation lead respectively to decreased mRNA synthesis and mRNA decay. Interestingly, impairment of RNA transcription also affected transcript stability, while RNA transcription was affected

by impairment of the mRNA degradation machinery (Sun et al., 2012).

The previous approaches involved comparing a transcript's nascent expression to its abundance in the mature RNA to calculate degradation rates. This can also be achieved by pulse-labeling the nascent RNA for a given time and chasing the labeling agent. By comparing the abundance of a transcript immediately after labeling to its abundance after a period of time has passed, one can calculate how much degradation took place within that period. This approach demonstrated that regulatory non-coding RNA and mRNA had a shorter half-life than housekeeping RNAs (Tani et al., 2012).

1.5 Treatments used to explore transcription

Chapter II explores the steady-state RNA synthesis and stability in several different cell lines. While that's very informative, certain questions about transcription cannot be answered unless the cell is perturbed. This allows one, for example, to study treatment-induced changes in synthesis and stability. Furthermore, well planned transcription disruptions can be used to learn details about specific events in the transcription process, such as the elongation rate of RNAPII. This section provides information on the treatments that were used to interfere with transcript synthesis and stability in chapters III to VI.

1.5.1 Tumor Necrosis Factor (TNF)

TNF was originally discovered when it was noticed that a factor released by activated macrophages had cytotoxic effects in neoplastic cell lines (Carswell et al., 1975). Through decades of studying, it has been determined that TNF (and other cytokines in the TNF superfamily) are some of the most important molecules involved in the pro-inflammatory response (Aggarwal et al., 2012). TNF's cellular response

is dependent on binding to one out of two possible type I transmembrane proteins, TNFRI or TNFRII (Hehlgans and Pfeffer, 2005). While a lot was previously known about the effect of TNF on steady-state RNA (Tian et al., 2005), the regulatory mechanism responsible for the changes in RNA expression were not well understood. On the other hand, it was known that transcript stability played an important role in the regulation of the inflammatory response (Hao and Baltimore, 2009; Anderson, 2010).

In Chapter III we explore the TNF induced effect on RNA synthesis and stability. Cells were exposed to TNF either prior to or after Bru labeling. The exposure of cells to TNF prior to labeling was carried out in order to determine the TNF induced effects on gene expression. On the other hand, cells were exposed to TNF after labeling in order to determine the effects of TNF in gene stability. The observed results are reported in sections 3.3.7, 3.3.8 and 3.3.9. Our study was the first to elucidate the TNF-mediated acute inflammatory response effect of RNA synthesis and stability.

1.5.2 Ultraviolet Light

Exposure to UV light can lead to various molecular effects in a cell, many of which are harmful. One of the immediate effects of UV exposure are DNA lesions. The most common type of DNA damage caused by UV light are cyclobutane pyrimidine dimers and 6-4 photoproducts. These lesions occur on neighboring pyrimidines and are most likely to occur when these bases happen to be thymines. The number of lesions in the genome is dependent on the UV dosage and its distribution is more or less random (Friedberg et al., 2006). These UV-induced lesions function as a barrier for the movement of RNAPII. When a traveling RNAPII enzyme encounters a damage site it stalls and recruits the transcription-coupled repair complex to remove the

damage. Independently from transcription, the global genomic machinery identifies UV damage sites in the whole genome. (Tornaletti and Hanawalt, 1999)

We exposed cells to UVC radiation prior to bromouridine-labeling and sequencing and reported the results in Chapter IV. Since UV-induced lesions block transcription elongation but not transcription initiation, the sequencing signal is enriched around TSS. In section 4.3.1, we show that this leads to greater signal around gene promoters, which can be used to determine TSS usage in nascent RNA. Unexpectedly, UV treatment also leads to an increase in the signal around potential enhancers identified by the ENCODE consortium, as demonstrated in sections 4.3.7. The promoter and enhancer aspects of this project are brought together when we demonstrate that treatment induced changes in gene expression are correlated to changes in enhancer RNA expression in section 4.3.8.

1.5.3 Camptothecin

DNA metabolic processes, such as transcription or replication, require the separation of the two strands of the duplex, which leads to the creation of DNA supercoiling upstream and downstream from the polymerizing complex. DNA topoisomerases are responsible for releasing the tension in the overwound DNA molecule. They cleave one (or both) of the DNA's strand, assist on the its controlled rotation and religate the nicked DNA molecule (Wang, 2002). DNA topoisomerase I (Top1) is inhibited by the drug camptothecin. This drug stabilizes the binding of Top1 to DNA prior to the controlled rotation stage (Hsiang and Liu, 1988). This becomes a barrier for the polymerizing complex which caused the tension build up in the first place. If the enzyme involved was RNAPII, the inhibition of Top1 by camptothecin leads to an inhibition in elongation (Ljungman and Hanawalt, 1996). Similarly, CPT treatment can cause a replication fork to be blocked. Since CPT inhibits Top1 after the DNA

has been nicked, the replication machinery creates a double-strand DNA break due to the run-off from the CPT-induced single-strand DNA break. This initiates a stress response mediated by the kinase ATM, which activates the p53 response pathway (Pommier, 2006).

In Chapter V, we explored how camptothecin exposure affected RNA synthesis. This treatment is extremely interesting, since it not only induces transcription changes, but it also interferes with the transcription machinery. In sections 5.4.1 and 5.4.2, we discuss the effects of a one hour long exposure to camptothecin on transcription. Another interesting aspect of this drug is that its inhibition of Top1 is reversible. In order to understand the cell's recovery from camptothecin, the drug was washed out and transcription was measured in the next 15 and 30 minutes. An analysis of the recovery of gene expression after camptothecin exposure is given in sections 5.4.3 to 5.4.5.

1.5.4 5,6-dichlorobenzimidazole 1-β-D-ribofuranoside (DRB)

As discussed in section 1.3.3, promoter-proximal pausing is regulated by DSIF and NELF (Wada et al., 1998a; Yamaguchi et al., 1999). Phosphorylation of these factors is necessary for the transition from early transcription initiation into productive transcription initiation. This process is carried out by the CDK9 kinase subunit of P-TEFb (Shandilya and Roberts, 2012). The drug 5,6-dichlorobenzimidazole 1-β-D-ribofuranoside (DRB) inhibits the phosphorylation of DISF and NELF by binding to CDK9 in a reversible manner (Zhu et al., 1997).

Since gene expression is not synchronized in cell cultures, certain genomic processes are very hard to study. An example is measuring the speed at which RNAPII translocates along a gene while transcribing it. In Chapter VI we exposed cells to DRB for one hour in order to inhibit the transition of RNAPII from initiation to

elongation transcription elongation. This was followed by DRB washout and immediate bromouridine labeling (Singh and Padgett, 2009) during 10 minutes. This allowed us to synchronize transcription initiation across the cells and capture the RNA produced within those 10 minutes. In sections 6.3.1 and 6.3.2, we demonstrate how this data was used to measure genome-wide elongation rates in five cell lines. Furthermore, transcript elongation rates were positively correlated across these cell lines. In sections 6.3.3 to 6.3.6 we attempt to identify genomic features that are associated with elongation rates in order to better understand factors that influence transcription elongation.

1.6 Bioinformatics challenges

In chapters II to VI, we use nascent RNA sequencing to explore transcription and RNA stability. The nature of the data created by nascent RNA sequencing techniques is very similar to that of the more commonly used RNA-seq (Nagalakshmi et al., 2008). The RNA populations assayed by these techniques, however, can be very different. This allows one to answer questions with nascent RNA sequencing which would be unanswerable using RNA-seq. In order to carry out such novel research it is necessary to develop computational tools specific for the questions being asked. In this section, I will expose some of these challenges and explore how they were approached.

1.6.1 Genomic read mapping

Answering the questions raised in this thesis relied heavily on the use of computational tools. Before tools could be developed to answer specific questions, it was necessary to create an analysis pipeline to carry out the initial and general steps to which every sample was subjected. The nascent RNA libraries were sequenced

using Illumina HiSeq 2000 machines at the University of Michigan Sequencing Core and downloaded into the University of Michigan Molecular and Behavioral Neuroscience Institute computing cluster. A full description of the basic pipeline is given in chapter II. The first step is to confirm the sequencing quality of the individual samples using the software fastQC (Andrews, 2010). If quality control standards are matched, the sample reads were initially mapped against the human ribosomal DNA complete repeating unit (RefSeq ID U13369.1) using Bowtie (Langmead et al., 2009). The reads that did not map to the ribosomal DNA were mapped against the human genome (hg19 assembly) using TopHat (Trapnell et al., 2009). In both cases, only reads that map uniquely, with up to two read segment mismatches, are accepted. Several different analysis are carried out based upon the mapping results. Most of these analysis rely greatly on standard bioinformatics and statistics tools such as Samtools (Li et al., 2009), BEDTools (Quinlan and Hall, 2010) and R (R Development Core Team, 2011).

1.6.2 RNA synthesis and stability measurements

RNA sequencing technologies are used to determine the abundance of RNA molecules in a given RNA population in the cell. In very basic terms, the RNA population of interest is isolated and its complementary DNA is sequenced and mapped to the genome. The number of reads mapped to a region of the genome is correlated to the amount of RNA transcripts that contain such sequence. Therefore, these techniques allow one to quantify RNA abundance in the cell. In order to normalize expression values to genomic feature length and library size, the reads per kilobase per million mapped reads (RPKM) expression metric used is usually used (Mortazavi et al., 2008). In RNA sequencing techniques that focus on mature RNA, most of the signal accumulates in exons. Therefore, the gene annotation used is limited to the exons

and only reads overlapping the exons are used to calculate transcription values. Due to the presence of intronic signal in nascent RNA, we used the signal distributed through the whole gene to determine synthesis values.

In order to calculate transcript stability, it was first necessary to measure transcript abundance after RNA processing and degradation took place. In our studies we allowed six hours to pass after the initial bromouridine labeling. During this period, most intronic signal was removed and degraded. Therefore, we measured RPKM values for the six hours old RNA based solely on the gene's exons annotation. In order to calculate transcript stability, a ratio between transcript abundance in six hour old RNA and RNA synthesis was calculated. This approach has the advantage of estimating RNA stability based solely on two samples. In previous work, half-life assessment was used to estimate RNA stability, which demanded a higher number of samples (Rabani et al., 2011; Tani et al., 2012).

1.6.3 *De novo* discovery of transcription units

One of the greatest characteristics of high-throughput sequencing techniques is that it allows observing genomic events in a non targeted way. Such approaches can be used to discover splice variants, binding sites for transcription factors, 3 dimensional chromatin interactions, and many other genome-wide processes (Wang et al., 2009; Park, 2009; Fullwood et al., 2009). In our nascent transcription studies, we observed that RNA synthesis was not limited to sites where genes had been previously annotated. In fact, we noticed that large transcription units existed in regions that would previously have been considered gene deserts. In order to identify such transcription units we used a Hidden Markov Model (HMM) (described in detail in chapter II). The nascent RNA expression values were used to organize the genome into segments with qualitatively different expression. This enabled us to identify

RNA producing genomic regions lacking gene annotation.

1.6.4 Using UV-induced signal redistribution to identify active TSS and putative enhancers

The genomic segmentation based on nascent RNA expression approach described in section 1.6.3 is extremely useful in classifying regions according to gene expression values. This signal becomes confounded, however, in genomic regions where overlapping transcription units are active. For example, several genes present multiple transcript isoforms which might share different portions of the same open reading frame. Understanding what percentage of the transcription signal is derived from which isoform is extremely challenging. In RNA-seq this problem is usually approached by analyzing the number of reads overlapping transcript-specific splice junctions with software such as Cufflinks (Trapnell et al., 2010). Since nascent RNA is mostly unspliced, such approaches would not be successful. By exposing cell cultures to UV light prior to bromouridine labeling, we are able to redistribute the sequencing signal around active TSS (see section 1.5.2). In chapter IV we describe how this signal accumulation can be used to identify which TSS are active. We used the widely encompassing ENSEMBL's gene annotation in order to assay as many known TSS as possible (Flicek et al., 2013). Basically, a two state HMM was used to identify a low expression pre-TSS state and a post-active TSS high expression state. When this transition happened within 500 base pairs of an annotated TSS, we considered it to be actively transcribing.

As discussed in section 1.6.3, due to the non targeted nature of sequencing techniques, focusing solely on annotated TSS can be very limiting. By comparing regular nascent signal with UV-irradiated nascent RNA signal, it is possible to recognize genomic regions where reads accumulated post irradiation (see chapter IV). Basically,

a two state HMM is used to classify the genome into UV repressed and UV enhanced regions. The emission probabilities of UV repressed regions are based upon the expression signal observed 20kb downstream from the TSS of expressed genes. Most UV enhanced regions overlapped gene's promoter regions. The signal around a significant proportion of UV enhanced regions, however, did not resemble gene expression. Further investigation demonstrated that these sites were potentially eRNA producing enhancers. These results indicated that UV irradiation prior to bromouridine labeling lead to an increase in the eRNA signal, similarly to how it enhanced TSS signal.

1.6.5 Measuring RNAPII elongation rate

Transcription elongation is a very important albeit poorly studied step in the transcription process (see section 1.3.4). An important aspect of transcription elongation is the rate at which RNAPII moves along the DNA molecule during transcription, which is called the elongation rate. Several studies have attempted to measure the elongating speed of RNAPII (Darzacq et al., 2007a; Wada et al., 2009; Singh and Padgett, 2009; Danko et al., 2013), but these studies were carried out in a small number of genes. In chapter VI, we arrested RNAPII at promoters for one hour. After this, we allowed transcription elongation and bromouridine labeling to happen during ten minutes. The goal of this analysis was to determine how far along the gene RNAPII had moved in the 10 minutes of labeling post DRB inhibition. Identifying the position of RNAPII on the genome is very challenging, therefore we used the binned expression signal as an indicator of the presence of transcribing RNAPII. The signal was strongest close to the promoter and steadily decreased until reaching background levels. We measured the elongation rate by identifying where the signal reverted to background levels. This was carried out using a three-state HMM. The

model is described on section 6.5.5 of the thesis. Basically, it was used to identify a region upstream from the promoter (state A), the transcription wave (state B) and the region downstream from the elongation wave (state C). The emission probabilities were calculated based on regions that displayed the signal of each state and the transition probability was set to 10^{-5}. The Viterbi algorithm was used to estimate the most likely states in each genomic position. These algorithms were used as implemented in the R package msm(Jackson, 2011).

1.6.6 Clustering of transcripts according to elongation rate

We used the technique described above to measure the elongation rate in five different cell lines. We observed a positive correlation between the transcript's elongation rate across every pair-wise comparison made between these cell lines. This prompted us to look for groups of transcripts that displayed similar elongation rate across cells (details on section 6.5.6). The Euclidean distance of elongation rates was used as the metric for the clustering. Due to the large size of the data set, we used the partitioning around medoids (PAM) algorithm (implemented in(Maechler et al., 2013)) to organize the genes into a predefined number of $k = 3$ clusters. Transcript's elongation rate across cell lines within clusters was very similar. This analysis identified groups of transcripts with high, intermediate and low elongation rates across cell lines.

1.6.7 Correlation between elongation rate and gene features

Due to the similarities in elongation rate observed across cell lines we postulated that the features present in the gene sequence could affect the elongation rate. The idea was that RNAPII would move faster or slower while transcribing certain sequences. If a relationship actually existed, the density of these features in the

genomic region covered by RNAPII in a transcript during labeling would be correlated to the elongation rate calculated for said transcript. This analysis was made more challenging due to the non-random distribution of certain features through the genome. Some features are very frequent in the proximity of promoters, so their density is high for any transcript with small elongation rate. Therefore, a regular regression analysis would not effectively test the correlation between these variables. In order to test the correlation between elongation rate and gene features we used a permutation analysis (details on section 6.5.8). Basically, a slope was calculated for the regression between gene feature density and elongation rate (o_i). Next, the transcript's elongation rates were randomized and the regression slope was calculated and stored (p_{in}). The previous step was repeated $n = 2,000$ times. A p-value was calculated by counting the percentage of the times that a value more extreme than o_i was observed within p_{in}. Lastly, since the correlation with several different gene features was tested, the p-values were corrected for multiple testing using the FDR method.

CHAPTER II

Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA

2.1 Abstract

Gene expression studies commonly examine total cellular RNA, which only provides information about its steady-state pool of RNA. It remains unclear whether differences in this steady-state difference reflects variable rates of transcription or RNA degradation. To specifically monitor RNA synthesis and degradation genome-wide, we developed Bru-Seq and BruChase-Seq. These assays are based on metabolic pulse-chase labeling of RNA using bromouridine (Bru). In Bru-Seq, recently labeled RNAs are sequenced to reveal spans of nascent transcription in the genome. In BruChase-Seq, cells are chased in uridine for different periods of time following Bru-labeling, allowing for the isolation of RNA populations of specific ages. Here we describe these methodologies in detail and highlight their usefulness in assessing RNA synthesis and stability as well as splicing kinetics with examples of specific genes from different human cell lines.

Official citation:
Paulsen, M.T., Veloso, A., Prasad, J., Bedi, K., Ljungman, E.A., Magnuson, B., Wilson, T.E., Ljungman, M. Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA. *Methods*, In Press. doi: 10.1016/j.ymeth.2013.08.015

2.2 Introduction

The steady-state level of a particular RNA in a cell is a balance between its rates of production and degradation. Production of RNA is regulated by epigenetic marks, transcription factors binding to enhancer and promoter elements and by release of RNA polymerases from transcription pause sites. Regulation of transcriptional elongation may also influence the output of RNA (Danko et al., 2013). Post-transcriptional regulation of RNA is mediated by the binding of miRNA or specific RNA-binding proteins to 3'-UTR sequences of selected transcripts to direct the recruitment of factors involved in RNA degradation (Houseley and Tollervey, 2009). Knowing the relative contribution of RNA synthesis and degradation to the steady-state level of particular transcripts is critical in order to better understand the mechanisms of regulation of these transcripts. Furthermore, when cell homeostasis is changed by environmental stimuli or stress and the steady-state levels of certain RNAs are altered, it would be of great interest to explore whether the ensuing gene expression changes were the result of altered RNA synthesis, stability or both.

A number of techniques have recently been developed to assess nascent RNA synthesis genome-wide. In global run-on and sequencing (GRO-Seq), nuclei from a cell sample are isolated and initiated RNA polymerases are allowed to "run-on" in vitro in the presence of bromouridine (Core et al., 2008). RNA polymerases that were arrested in vivo are released in vitro revealing which promoters harbored arrested RNA polymerases. GRO-Seq also allows for the detection of unstable RNAs such as promoter divergent transcripts since very little RNA degradation takes place in the in vitro run-on assay (Core et al., 2012). In native elongating transcript sequencing (NET-Seq), nascent RNA is isolated by immunoprecipitation of the RNA polymerase

II elongation complex followed by deep sequencing of the 3' ends of nascent transcripts associated with the RNA polymerases (Churchman and Weissman, 2011). This technique allows for nucleotide-level resolution of nascent transcription and has revealed that RNA polymerase II frequently pauses and backtracks when encountering nucleosomes in the bodies of genes (Churchman and Weissman, 2011). Nascent-Seq is based on the isolation of chromatin-bound nascent RNA obtained from the lysis of cells and washing of cell nuclei with NUN buffer consisting of high concentrations of NaCl, urea and NP-40 (Khodor et al., 2011). This technique has been used to monitor the efficiency of intron splicing and has provided evidence that not all splicing events occur co-transcriptionally. A different approach to assess nascent transcription is through metabolic labeling of RNA with tagged ribonucleotides followed by isolation and analysis using microarrays or deep sequencing (Ohtsu et al., 2008; Rabani et al., 2011; Schwanhusser et al., 2011; Schwalb et al., 2012; Sun et al., 2012). This approach has been extended to also estimate the half-lives of transcripts by computationally comparing nascent and steady-state levels of RNA.

Bromouridine sequencing (Bru-Seq) and bromouridine-chase sequencing (BruChase-Seq) are based on the metabolic pulse-chase labeling of nascent RNA with bromouridine. Bromouridine has been used to label steady state RNA (Tani et al., 2012) and nascent RNA (Haider et al., 1997; Ohtsu et al., 2008) both in vitro and in cells (Core et al., 2008). While other ribonucleotide analogs, such as 4-thiouridine (4sU) and ethynyluridine (EU), can be used to specifically label and isolate nascent RNA, bromouridine is less toxic to cells than these other analogs (Tani et al., 2012). Furthermore, the low cost of bromouridine and the availability of excellent anti-BrdU antibodies make bromouridine labeling of nascent RNA an attractive approach to study transcriptional and post-transcriptional regulation.

Following labeling, Bru-containing RNA is specifically captured using anti-BrdU antibodies conjugated to magnetic beads. cDNA libraries are then produced from the isolated Bru-RNA and subjected to deep sequencing (Paulsen et al., 2013b). By chasing Bru-labeled cells with uridine for different periods of time, RNA populations of defined ages can be isolated and analyzed. This allows for the estimation of the relative stability of all transcripts and splicing kinetics of all introns. We recently used these techniques to obtain signatures of the TNF-induced acute inflammatory response in human fibroblasts and found a complex pattern of altered synthesis and/or stability of specific RNAs (Paulsen et al., 2013b). We also found interesting patterns of synthesis, stability and splicing in untreated cells suggesting that steady-state RNA levels are controlled by intricate transcriptional and post-transcriptional regulation.

Here we describe Bru-Seq and BruChase-Seq in detail and show examples of how the stability of transcripts vary in a cell type-specific manner. Furthermore, we show that BruChase-Seq can be used to predict nonsense and frameshift mutations in genes by revealing increased mRNA turnover rates. Finally, using segmentation analysis of nascent transcription spans we show how Bru-Seq can detect unannotated, long non-coding RNAs (lncRNA) with a highly cell type-specific expression pattern.

2.3 Description of Methods

The Bru-Seq and BruChase-Seq techniques were recently described (Paulsen et al., 2013b). We will here provide a more detailed description of the materials and procedures involved in the different steps of these techniques.

2.3.1 Materials

Buffer	Composition
RPMI growth medium	PMI 1640, 10% FBS, 100 U/ml penicillin, 100 U/ml streptomycin
DMEM growth medium	DMEM, 10% FBS, 100 U/ml penicillin, 100 U/ml streptomycin
MEM growth medium	Minimal Essential Medium, 10% FBS, 1X MEM Amino Acids, 1X Non-Essential Amino Acids, 2 mM L-glutamine, 1X antibiotic-antimycotic, 1X MEM vitamin mixture, 0.15% (w/v) sodium bicarbonate
6x gel loading buffer	10mM Tris, pH 7.6, 60% glycerol, 60 mM EDTA, 0.03% bromophenol blue

Material	Supplier, catalog number
(-)-5-Bromouridine	Sigma-Aldrich, 850187
Uridine	Sigma, U3750
Trizol Reagent	Invitrogen, 15596-018
Chloroform	Fisher, BP1145
Isopropanol	Sigma-Aldrich, 190764
Diethyl pyrocarbonate (DEPC)	Sigma, D5758
Bovine Serum Albumin (BSA)	Roche, 03116999001
Dynabeads Goat anti-Mouse IgG	Invitrogen, 110.33
Mouse anti-BrdU	BD Pharmingen, 555627

Material	Supplier, catalog number
RNaseOUT, Ribonuclease Inhibitor	Invitrogen, 10777-019
Superscript II	Invitrogen, 18064-014
Random Primers ($3\mu g/\mu l$)	Invitrogen, 48190-011
100mM dNTP set	Invitrogen, 10297-018
ActinomycinD	Sigma, A9415
AmPure RNAclean beads	Fisher, APN000494
10X NEBuffer 2	New England Biolabs, B7002S
dUTP	Roche, 11934554001
RNase H	Invitrogen, 18021-014
DNA Polymerase I	Invitrogen, 18010-017
AmPure XP beads	Fisher, NC9933872
TruSeq RNA Preparation Kit	Illumina, RS-122-2001
NuSieve 3:1 agarose	Lonza, 50090
10X TAE Buffer	Lonza, 50844
50bp ladder	Invitrogen, 10416-014
Gel Excision Tips	The Gel Company, PKB6.5-R
QIAEX II Gel Extraction Kit	Qiagen, 20021
USER enzyme	New England Biolabs, M5505L

2.3.2 Procedures

2.3.2.1 Cell culturing

1. Grow cells in appropriate growth medium. For this study, RPMI (BxPC3), DMEM (Panc1, MiaPaCa2, HeLa) and MEM (NF) were used.

2. Follow normal cell culture protocols to expand cells. For most cell lines, we recommend using 2 to 3 10-cm plates, or a minimum of 4×10^6 cells per sample.

3. Cells are grown to approximately 80% confluency before the addition of bromouridine.

2.3.2.2 Bromouridine labeling

- Make a stock solution of 50 mM Bromouridine in PBS
- Make a stock solution of 1 M uridine in PBS for chase (stability analysis)
- Use conditioned media for all treatments

1. Remove 3-4 ml of media from each plate of cells to a clean tube and add BrU to a final concentration of 2 mM. Discard remaining media from plate, or save to use for a uridine chase.

2. Add back BrU-containing media to plate and incubate at 37°C for 30 min.

3. If doing a chase, after the 30 min incubation, rinse plate twice with PBS, then add back saved media containing 20 mM uridine and incubate for desired time period (6 hours may be an appropriate time to start with).

4. To collect cells (pooling plates as necessary), either add Trizol directly to the plate to lyse cells, or trypsinize cells, spin to pellet and resuspend in 3-5 ml Trizol. Vortex until no cell pellet is visible. Store samples at −80°C if not

isolating RNA immediately. We recommend collecting the cells in 14ml round-bottom centrifuge tubes (e.g. BD 352059).

2.3.2.3 Isolation of RNA

1. To each Trizol-lysed sample, add 0.2 ml chloroform per 1ml of Trizol used initially. Cap tube and shake vigorously for 5-10 seconds. Remove cap, cover tube with parafilm and centrifuge at 4°C for 15 min at 12,000g in a Sorvall RC5C floor centrifuge with SS-34 rotor (or equivalent).

2. Transfer the upper aqueous layer to a new round-bottom 14 ml tube and add 0.5 ml isopropanol per 1 ml of Trizol used initially. Cover tube with parafilm and mix gently. Incubate at room temperature for 10 min before centrifuging at 4°C for 10 min at 12,000g to pellet RNA.

3. Remove supernatant and wash pellet by adding 1 ml of 75% ethanol per 1ml of Trizol used initially. Cover tube with parafilm and centrifuge at 4°C for 5 min at 7,500g.

4. Remove supernatant and invert tube to allow pellet to dry slightly. Resuspend the pellet in 200μl DEPC-water and incubate at 55°C for 10 min to ensure RNA is fully dissolved. Store RNA at -80°C unless immediately proceeding to isolation of Bru-RNA.

2.3.2.4 Preparation of magnetic beads conjugated with anti-BrdU antibodies

1. Transfer 50μl of anti-mouse IgG magnetic Dynabeads (Invitogen) per sample to a 1.5 ml microfuge tube. Capture beads with a magnetic stand (Novagen) and aspirate storage buffer.

2. Add 200μl 0.1% BSA in DEPC-PBS, flick the tube to resuspend beads, capture

beads on the magnetic stand and aspirate supernatant. Repeat 2 more times for a total of 3 washes. After the final wash, resuspend each bead pellet in 200μl 0.1% BSA in DEPC-PBS and add 0.5μl RNaseOUT.

3. To each tube, add 4μl (2μg) anti-BrdU antibody and 0.5μl (20U) RNaseOUT. Incubate with gentle rotation for 1 hour at room temperature.

4. Wash beads 3 times with 200μl 0.1% BSA in DEPC-PBS as detailed above. After the final wash, resuspend conjugated beads in 200μl 0.1% BSA in DEPC-PBS and add 0.5μl RNaseOUT.

2.3.2.5 Isolation of Bru-labeled RNA

1. Heat isolated RNA in an 80°C heat block for 10 min, then immediately put samples on ice.

2. Remove 90% of the sample (180μl) and add to prepared beads. Flick tube to ensure sample is mixed well and place on rotator for 1 hour at room temperature.

3. Wash beads with 0.1% BSA in DEPC-PBS for 5 minutes on the rotator. Do 2 additional brief washes with 0.1% BSA in DEPC-PBS, making sure to completely remove the final wash.

4. Resuspend the bead pellet in 40μl DEPC-water and incubate for 10 min in a 95°C heat block to elute Bru-RNA from the beads.

5. Centrifuge tubes briefly, then capture beads in the magnetic stand.

6. Remove the supernatant to a clean 1.5 ml microfuge tube, quantitate Bru-RNA concentrations (Nanodrop, Thermo Scientific), and store at −80°C if not using immediately for library preparation.

2.3.2.6 cDNA library preparation

- Start with at least 250 ng Bru-RNA
- Unless otherwise stated, use a Thermomixer R (Eppendorf) for all incubations.

Fragment mRNA

1. Pre-mix (per sample):
   (a) $8\mu l$ 5x First-strand buffer (comes with Superscript II)
   (b) $1\mu l$ Random primer ($3\mu g/\mu l$)
2. Add $9\mu l$ of pre-mix to each PCR tube
3. Add $16\mu l$ RNA to each tube
4. Incubate in PCR machine at 85°C for 10min. Cool down to 4°C

Synthesize First Strand cDNA (for strand specificity)

1. Mix the following reagents (per sample):
   (a) $4.0\mu l$ 100 mM DTT (comes with Superscript II)
   (b) $0.8\mu l$ 25 mM dNTP
   (c) $0.5\mu l$ RNaseOUT
   (d) $0.8\mu l$ ActinomycinD ($2.5\mu g/\mu l$ stock)
   (e) $6.9\mu l$ ddH2O
   (f) $2.0\mu l$ Superscript II
2. Add $15\mu l$ mixture to each $25\mu l$ RNA sample.
3. Incubate samples on thermal cycler using the following program:

(a) 25°C for 10 minutes

(b) 42°C for 50 minutes

(c) 70°C for 15 minutes

(d) Hold at 4°C

Purify First Strand DNA with AMPure RNAclean beads

1. Mix 40μl cDNA mixture with 72μl RNAclean beads.
2. Bind at room temp for at least 10 min (with 500 rpm shaking).
3. Wash beads twice with 80% ethanol. Spin briefly and remove any additional ethanol.
4. Dry beads at 37°C for $3min$.
5. Add 42μl of 5 mM Tris, pH 8.0, mix well.
6. Incubate at 28°C for 10-15 min (with shaking).
7. Capture beads and transfer 40μl supernatant to a new tube.

Synthesize Second Strand cDNA

1. Mix the following reagents:
   (a) 20μl 10X NEBuffer 2
   (b) 1.2μl 25 mM dG+dA+dU+dC mix
   (c) 35.3μl ddH2O
   (d) 1μl RNase H
   (e) 5μl DNA polymerase I
2. Add 60μl mixture to each 40μl First strand sample.

3. Incubate samples on thermal cycler at 16°C for 2.5 hours

Clean Up with AMPure beads

1. Vortex AMPure beads and add 150μl to each 100μl sample.
2. Pipette up and down to mix thoroughly.
3. Incubate at room temp with shaking for 10-15 min.
4. Capture beads for 5 min.
5. Remove and discard supernatant from each sample.
6. With the tubes still in the stand, add 200μl freshly prepared 80% ethanol to each sample without disturbing the beads
7. Incubate at room temp for 30 sec, then remove and discard supernatant from each sample.
8. Repeat steps 6 & 7 for a total of two ethanol washes.
9. Spin samples briefly and remove any remaining ethanol
10. Incubate samples at 37°C until dry.
11. Add 62μl Resuspension Buffer (from TruSeq kit) to each sample
12. Pipette up and down to mix thoroughly.
13. Incubate samples at room temp (with shaking) for 10-15 min.
14. Place samples on magnetic stand for 5 min.
15. Transfer 60μl of the supernatant (ds cDNA) to a new tube.

*** Can stop here and store samples at -20°C for up to seven days

From this step (End Repair) forward, the reagents will be from Illumina's TruSeq Kit Perform End Repair

- Thaw End Repair Mix, Resuspension Buffer at room temperature.
- Make sure AMPure beads are at room temperature.
- Pre-heat Thermomixer to 30°C

1. Add 40μl End Repair Mix to each 60μl sample.
2. Adjust pipette to 100μl and pipette up and down to mix thoroughly.
3. Incubate samples in Thermomixer (no shaking) at 30°C for 30 min.

Clean Up with AMPure beads

1. Vortex AMPure beads and add 160μl to each sample.
2. Pipette up and down to mix thoroughly.
3. Incubate at room temp (with shaking) for 10-15 min.
4. Capture beads for 5 min or until liquid appears clear.
5. Remove and discard supernatant from each sample.
6. With the tubes still in the stand, add 200μl freshly prepared 80% ethanol to each sample without disturbing the beads.
7. Incubate at room temp for 30 sec, then remove and discard supernatant from sample.
8. Repeat steps 7 & 8 for a total of two ethanol washes.
9. Spin samples briefly and remove any remaining ethanol.

10. Incubate samples at 37°C until dry.
11. Resuspend the dried pellet with 20μl Resuspension Buffer.
12. Pipette up and down to mix thoroughly.
13. Incubate samples at room temp (with shaking) for 10-15 min.
14. Place samples on magnetic stand for 5 min, or until liquid appears clear
15. Transfer 17.5μl of the supernatant to a new tube.

*** Can stop here and store samples at −20°C for up to seven days

Adenylate 3' Ends

- Thaw Resuspension Buffer and A-Tailing Mix at room temp
- Pre-heat Thermomixer to 37°C

1. Add 12.5μl A-Tailing Mix to each sample
2. Pipette up and down to mix thoroughly.
3. Incubate samples at 37°C for 30 min (no shaking).

Ligate Adaptors

- Thaw RNA Adaptor Index tubes, Stop Ligation Buffer, and Resuspension Buffer at room temp.
- Make sure AMPure beads are at room temp
- Pre-heat Thermomixer to 30°C

1. Add 2.5μl Resuspension Buffer to each sample.

2. Add 2.5μl Ligation Mix to each sample. (Remove Ligation Mix from -20°C just before using and return to -20°C immediately after using).
3. Add 2.5μl desired RNA Adaptor Index to appropriate sample.
4. Adjust pipette to 37.5μl and pipette up and down to mix thoroughly.
5. Incubate samples in Thermomixer at 30°C for 10 min (no shaking).
6. Add 5μl Stop Ligation Buffer to each sample.
7. Adjust pipette to 42.5μl and pipette up and down to mix thoroughly.

Clean Up with AMPure beads

1. Vortex AMPure beads and add 65μl to each sample.
2. Pipette up and down to mix thoroughly.
3. Incubate at room temp for 15 min.
4. Capture beads for 5 min or until liquid appears clear.
5. Remove and discard supernatant from each sample.
6. With the tubes still in the stand, add 200μl freshly prepared 80% ethanol to each sample without disturbing the beads.
7. Incubate at room temp for 30 sec, then remove and discard supernatant from sample.
8. Repeat steps 6 & 7 for a total of two ethanol washes.
9. Spin samples briefly and remove any remaining ethanol
10. Incubate samples at 37°C until dry.

11. Resuspend the dried pellet with 32μl 5 mM Tris.
12. Pipette up and down to mix thoroughly.
13. Incubate samples at room temp (with shaking) for 10-15 min.
14. Place samples on magnetic stand for 5 min, or until liquid appears clear
15. Transfer 30μl of the supernatant to a new tube.

*** Can stop here and store samples at -20°C for up to seven days

Size Selection by Agarose gel electrophoresis

- Cast 3% gel using NuSieve 3:1 agarose
- Remove buffer from wells before loading
- Load order: ladder, sample, ladder, sample, ladder, sample, etc.
- Run gel in 1XTAE and do not cover gel with buffer

1. Add 5μl 6x gel loading buffer to each sample
2. Use 50 bp ladder
3. Load the gel and run at 65 V for 1hour 40min
4. Rinse gel with distilled water
5. Excise gel slices in the 300 bp region using a gel excision tip with a 1000μl pipettor. Cut out backup gel slice as well of a 350 bp size.
6. Purify the gel slices with QIAEXII kit as follows:
   (a) Add 900μl QX buffer and 10μl QIAEX II suspension beads and mix well.
   (b) Incubate at 40°C with shaking for 15 min

(c) Spin 13,000 rpm for 30 sec. Remove supernatant

(d) Add 500μl QX, vortex, spin at 13,000 rpm for 30 sec

(e) Remove supernatant, add 500μl PE buffer, spin at 13,000 rpm for 30 sec

(f) Repeat PE wash

(g) Remove supernatant, spin again at 13,000 rpm

(h) Remove supernatant, dry beads at 37°C until they turn white.

7. Add 22μl Resuspension Buffer to elute DNA. Mix well and incubate at room temperature (with shaking) for 10-15 min.

8. Transfer 20μl of the supernatant to a PCR tube.

*** Can stop here and store samples at -20°C for up to seven days

Uridine Digestion/Enrich DNA Fragments

- Thaw PCR Master Mix and PCR Primer Cocktail at room temp and spin briefly.
- Make sure AMPure beads are at room temperature.

1. Mix the following reagents (per sample):

(a) 25.0μl PCR Master Mix

(b) 5.0μl PCR Primer Cocktail

(c) 1.0μl USER enzyme

2. Add 31μl to each sample.

3. Pipette up and down to mix thoroughly

4. Incubate samples on thermal cycler using the following program:

(a) 37°C for 15 min (uridine digestion)

(b) 98°C for 30 sec

(c) 15 cycles of:

i 98°C for 10 sec

ii 60°C for 30 sec

iii 72°C for 30 sec

(d) 72°C for 5 min

(e) Hold at 10°C

Clean Up with AMPure beads

1. Vortex AMPure beads and add 50μl to each sample.
2. Pipette up and down to mix thoroughly.
3. Incubate at room temp with shaking for 10-15 min.
4. Capture beads for 5 min or until liquid appears clear.
5. Remove and discard supernatant from each sample.
6. With the tubes still in the stand, add 200μl freshly prepared 80% ethanol to each sample without disturbing the beads.
7. Incubate at room temp for 30 sec, then remove and discard supernatant from sample.
8. Repeat steps 6 & 7 for a total of two ethanol washes.
9. Spin samples briefly and remove any remaining ethanol
10. Incubate samples at 37°C until dry.
11. Resuspend the dried pellet with 27μl 5mM Tris Buffer.

12. Pipette up and down to mix thoroughly.
13. Incubate at room temp with shaking for 10-15 min.
14. Place samples on magnetic stand for 5 min, or until liquid appears clear
15. Transfer 25μl of the supernatant to a new PCR tube.

Validation

1. Use 3μl of each library to run on a thin 1.5% agarose gel to ensure there is a single band running around 300 bp
2. Quantitate libraries using Nanodrop. Set 20μl of sample aside and save. The rest of the samples are now ready for sequencing.

2.3.3 Deep sequencing

Sequencing can be performed using any preferred platform. We use Illumina HiSeq 2000 via the University of Michigan Sequencing Core. We also take advantage of cost-saving associated with sample indexing. Acceptable results can often be obtained for most expressed genes when the reads (40 million at a time, in our case) are distributed across multiple samples.

2.3.4 Data analysis pipeline

The conceptual bioinformatics approaches used in Bru-Seq and BruChase-Seq were recently described (Paulsen et al., 2013b). Our data analysis pipeline, which uses common bioinformatics tools for sequence read analysis (section 2.4.1) as well as custom scripts, is implemented using the q pipeline manager (`http://sourceforge.net/projects/q-ppln-mngr/`).

2.3.4.1 Major programs used

Program	Version
TopHat (Trapnell et al., 2009)	v1.4.1
Bowtie (Langmead et al., 2009)	v0.12.8
BEDTools (Quinlan and Hall, 2010)	v2.16.2
Samtools (Li et al., 2009)	v0.1.18
R (R Development Core Team, 2011)	v2.15.1
DESeq (Anders and Huber, 2010)	v1.4.1

2.3.4.2 Read mapping (q master map)

1. Map reads to the human ribosomal DNA complete repeating unit (U13369.1) using Bowtie; keep rRNA read counts and non-rRNA read sequences.

2. Map non-rRNA reads to the human reference genome assembly hg19/GRCh37, or other appropriate genome, using TopHat.

   (a) Keep only reads that map uniquely, with up to two read segment mismatches.

   (b) Reads are allowed to split between exons in RefSeq or another preferred transcript annotation, but de novo splice junction calling is not performed since nascent RNA reads are mainly intronic.

   (c) Duplicate reads are maintained and expected in mature RNA samples where reads cluster in exons.

2.3.4.3 Genome annotation

1. In preparation for counting, condense the RefSeq transcript isoforms of genes into one BED file of non-redundant intron and exon spans, using create_

transcriptome_map.pl (`http://tewlab.path.med.umich.edu/software/utilities/utilities.html`) or another utility, so that genome bases will have only one assigned identity.

(a) When isoforms conflict, give priority to annotation as an exon to prevent a stable exon from being annotated as an intron.

(b) Overlapping regions of different genes are termed ambiguous and ignored when determining the expression level of the involved genes.

2.3.4.4 Expression scoring (q master map)

1. Determine the strand-specific coverage over each genome base so that a read might be fractionally attributed to different exons, bins or other features.

   (a) Count the number of reads in a given orientation overlapping each base, using BEDTools.

   (b) Divide by the length of the sequenced reads.

2. Sum the base coverages across a given feature to determine its read coverage.

   (a) For gene expression in Bru-Seq samples, calculate the RPKM using all introns and exons.

   (b) For gene expression in BruChase-Seq samples, calculate the RPKM using all, and only, exons.

3. Similarly sum the base coverages and calculate RPKM for each 1 Kb genome bin, or other desired bin size, in preparation for segmentation.

2.3.4.5 Combining replicates (q master merge)

1. Sum the fractional base coverages and bin coverages over all replicate samples and recalculate feature and bin RPKM as in 2.4.4.

2.3.4.6 Genome segmentation (q master segment)

1. Normalize the 1 kb genome bins by discarding unmappable bins and dividing remaining bin RPKM values by the fractional mappability, determined using extractKmers.pl or another utility, to prevent unmappable regions from breaking contiguous transcription units.

2. Apply wavelet smoothing to the normalized bin RPKM using smooth.pl (`http://sourceforge.net/projects/smooth-stream/`) or another utility.

3. Establish the emission probabilities of a hidden Markov model:

   (a) Score the bins by rounding each into one of 17 logarithmically distributed RPKM input states.

   (b) Score annotated genes by rounding each into one of 10 logarithmically distributed RPKM output states.

   (c) Assign emission probabilities as the frequency of input bin states observed over all gene output states. See Supplementary Figure S2.1.

4. Set the transition probability to 0.005 for all bins.

5. Solve the model using the Viterbi algorithm using segment.pl (`http://sourceforge.net/projects/segment-stream/`) or another utility to establish the most likely bin expression states.

6. Fuse adjacent bins of the same state into genome segments of sustained contiguous expression.

2.3.4.7 Rank synthesis and stability (q master assemble)

1. Score relative synthesis by ranking the RPKM of genes, segments, or other features obtained from nascent RNA collected immediately after Bru labeling.

2. Score relative stability by ranking the ratio of the RPKM of features obtained from aged RNA samples over paired nascent RNA samples.

3. Score splicing extent as intron retention, i.e. the RPKM of a specific intron divided by the RPKM of all exons of the same gene.

2.3.4.8 Inter-sample comparisons (q master compare)

1. Apply the DESeq R package to compare the replicates that gave rise to one merged sample to the replicates of a different merged sample of the same type (nascent or aged).

   (a) Use genes and other annotated feature as is.

   (b) Split hidden Markov segments at all inter-segment boundaries encountered in either sample to provide the complete set of potentially divergent transcription segments.

2. Compare Bru-Seq samples to explore differences in nascent RNA synthesis for genes, individual exons, or other features.

3. Compare BruChase-Seq samples to explore differences in RNA stability, but only when both samples were split from the same Bru-labeled cell stock prior to a manipulation that might affect transcript stability, so that the input nascent RNA is identical.

2.4 Results

2.4.1 BruChase-Seq reveals cell type-specific regulation of RNA stability

RNA degradation is regulated by specific miRNA and RNA-binding proteins that bind to the 3'-UTR or internal sequences of mature transcripts. By comparing the amount of exonic RNA reads present 6 hours after Bru-labeling with the total

amount of reads for the entire gene directly after Bru-labeling, an estimation of the relative stability of each transcript can be made. To test whether RNA stability is transcript-specific or whether the stabilities of specific RNAs differ in a cell type-specific manner, we performed BruChase-Seq on a set of human cell lines. As can be seen in Figure 2.1a, the NFKB1 transcript showed robust stability in normal fibroblasts as determined by the relatively high exonic signal at 6 hours (red) compared to the nascent transcript level (blue). However, the exonic signal was found to be much less prominent at 6 hours in either HeLa or H146 cells (Fig. 2.1b&c). In contrast, the transcript of DPC2 exhibited a low relative level of exonic signal in human fibroblasts (Fig. 2.1d) while the relative exonic signal was much higher in K562 and H146 cell lines suggesting that this transcript is more stable in these cancer cell lines (Fig. 2.1e&f). These results demonstrate the usefulness of BruChase-Seq in monitoring the relative stability of transcripts and indicate that the stability of some transcripts varies in a cell type-specific manner.

2.4.2 Stability of the MYC transcript is elevated in some cancer cell lines

The expression of the oncoprotein MYC is frequently upregulated in human cancers. This upregulation is sometimes caused by amplification of the MYC gene and in the case of HeLa cells is due to integration and amplification of viral regulatory regions proximal to the MYC gene (Lazo et al., 1989; Macville et al., 1999). To test whether MYC transcripts may be regulated at the level of stability in cancer cells we used BruChase-Seq. The MYC transcript was found to be quite unstable in both human fibroblasts and HeLa cells as previously shown for many cell lines Dani et al. (1984) (Fig. 2.2a&b). However, in the pancreatic cancer cell lines BxPC3 and MiaPaCa2, the exonic reads at 6 hours were quite robust suggesting that the MYC transcript is more stable in these cell lines as compared to human fibroblasts

and HeLa cells. The BruChase-Seq technique thus revealed that the MYC transcript shows differential stability in different cell lines, suggesting that regulation of transcript stability may contribute to MYC overexpression in human tumors.

2.4.3 Nonsense and frame-shift mutated transcripts show low stabilities

Transcripts containing premature translation termination codons are targeted by the nonsense-mediated decay (NMD) pathway, removing these defective transcripts through RNA degradation during attempted translation (Brogna and Wen, 2009). To test whether BruChase-Seq can reveal a higher rate of degradation of transcripts from genes bearing nonsense mutations, we compared the relative stabilities of RB1 mRNA in human fibroblasts, where the RB1 gene is wild-type, and in H146 cells where it carries a nonsense mutation (CCLE, Broad Institute). The wild-type RB1 transcript had high exonic reads at 6 hours in human fibroblasts (Fig. 2.3a). However, the mutated RB1 transcript in H146 cells had a low level of exonic reads at 6 hours (Fig. 2.3b). Frameshift mutations can also lead to the formation of premature translation termination codons activating NMD (Pereira et al., 2006; Micale et al., 2009). The TP53 gene in H146 cells contains such a frameshift mutation Forbes et al. (2006). To explore whether it causes the TP53 transcripts to become unstable, we used BruChase-Seq comparing the exonic reads of TP53 in human fibroblasts and in H146 cells. The wild-type TP53 transcripts had a high level of exonic reads after a six-hour chase in the fibroblasts while the mutant TP53 transcript in H146 cells had much lower levels of exonic reads, demonstrating that it is much less stable (Fig. 2.3c&d). Using the BruChase-Seq approach, we have confirmed that the stabilities of mutant RB1 and TP53 transcripts in H146 cells are reduced, presumably through the activation of NMD.

2.4.4 Using BruChase seq to explore splicing kinetics

A variation of BruChase-Seq is to use different durations of uridine chase to allow RNAs of defined ages to be isolated and analyzed. This approach is very useful when exploring post-transcriptional processing of primary transcripts. Four different ages of RNA expressed from the CD44 gene are shown in Figure 2.4. At 0 hours (nascent RNA), the complete and predominantly unspliced primary transcript can be observed. After a 2-hour chase, most of the intronic signal has disappeared as a function of splicing and degradation, while the reads covering exons and the 3'-UTR are enhanced. One region of the CD44 transcript, however, is not processed with the same kinetics as other regions. This represents the so called "variable region" of the CD44 transcript, which is commonly omitted in the mature transcript of most cell types (Tölg et al., 1993). The variable region, notably, was still present after 4 hours, but by 6 hours it had nearly disappeared. This "retention" of the variable region over time may be due to slow splicing of the introns in this region, which would suggest that this process occurs post-transcriptionally. Alternatively, most of the transcripts may have their variable regions spliced out co-transcriptionally, but the transcripts retaining the variable region are subjected to accelerated degradation and are lost more quickly than the fully-spliced population. A third possibility is that the variable region is co-transcriptionally spliced but that this spliced intron possesses a non-coding function allowing it to escape degradation. Regardless of the mechanism, BruChase-seq allows determination of intron retention genome-wide and should provide new insights into the regulation of splicing in human cells.

2.4.5 Bru-Seq reveals cell type-specific expression of long, non-coding RNAs

The use of RNA-Seq technology has revealed a myriad of lncRNAs generated throughout the genome (Wang and Chang, 2011). The functions of these RNA species are mostly unknown. By emplying a hidden Markov model-based segmentation analysis, we were able to identify transcription units independently of prior gene annotation. We applied this method across five different cell types (human whole blood, normal human fibroblasts and the cancer cell lines K562, BxPC3 and Panc1) and identified numerous unannotated lncRNAs, . some of which were very large, spanning over 100 Kb in length (Fig. 2.5a). When compared across the five cell types, these lncRNAs exhibited distinctive cell type-specific expression patterns. The lncRNA shown in Figure 2.5a was only found to be expressed in BxPC3 cells while the lncRNA shown in Figure 2.5b was expressed only in whole blood cells. Human fibroblasts contained the least (15) and K562 cells contained the most (153) previously unannotated lncRNAs of the five cell types analyzed (Fig. 2.5c). Strikingly, most of the lncRNAs were uniquely expressed in each cell line and it appeared that the lncRNA transcripts were either highly expressed or not expressed at all (Fig. 2.5 a&b). We conclude that Bru-Seq can be used to identify non-annotated lncRNA genome-wide and our results suggest that these RNAs have a very strong cell type-specific expression pattern.

2.5 Conclusions

To better understand the underlying mechanisms regulating the steady state levels of RNA in cells, the contributions of both RNA synthesis and RNA degradation must be taken into account. To study both RNA synthesis and degradation in living cells, we developed the Bru-Seq and BruChase-Seq approaches based on metabolic

pulse-chase labeling of nascent RNA with bromouridine Paulsen et al. (2013b). We have here outlined these techniques in detail and have provided examples of their usefulness in assessing RNA synthesis and RNA stability for selected genes across multiple cell lines. The results obtained with these techniques suggest that transcript stability is differentially regulated in different cell types. For example, NFKB1 transcript were turned over faster in the cancer cell lines HeLa and H146 than in normal human fibroblasts while the DCP2 transcript is more stable in the cancer lines K562 and H146 than in the human fibroblasts. The MYC transcript was very unstable in human fibroblasts and HeLa cells but much more stable in the cancer cell lines BxPC3 and MiaPaCa2. BruChase-Seq also confirmed that mutations causing premature translation termination codons generate highly unstable transcripts, probably due to the activation of NMD to degrade these defective transcripts.

Comparing RNA populations of different ages, obtained from cells chased for different periods of time following bromouridine labeling, allows for the detailed exploration of post-transcriptional processing events such as splicing. As an example, BruChase-Seq revealed a much slower removal of the variable region of the CD44 compared to adjacent introns. The reason for this is not clear but it could be due to (i) slow splicing, (ii) a faster decay of transcripts containing a retained variable region or (iii) enhanced stability of the spliced intron, perhaps due to the presence of a putative functional sequence. We are currently exploring the regulation of splicing and intron retention genome-wide using BruChase-Seq.

Finally we showed examples of the power of Bru-Seq to identify nascent RNA transcripts outside of any prior annotation via segmentation analysis. We found that most unannotated lncRNA species were expressed in a very cell type-specific manner when comparing across five different cell lines. The functions of these lncRNAs and

how they are regulated are poorly understood but of great interest, and the Bru-Seq technique is ideally suited to identify and characterize expression of these RNA species in diverse cell types and during cellular processes such as differentiation and transformation.

2.6 Acknowledgments

We are grateful for the assistance by Manhong Dai and Fan Meng for administration and maintenance of the University of Michigan Molecular and Behavioral Neuroscience Institute (MBNI) computing cluster and by the personnel at the University of Michigan Sequencing Core. This work has been supported by funds from University of Michigan Bioinformatics Program, University of Michigan Biomedical Research Council, the Will and Jeanne Caldwell Endowed Research Fund of the University of Michigan Comprehensive Cancer Center, University of Michigan School of Public Health (NIEHS P30), Department of Defense, Uniting Against Lung Cancer, University of Michigan Nathan Shock Center, University of Michigan Office of the Vice President of Research, National Cancer Institute (5R21CA150100), National Institute of Environmental Sciences (1R21ES020946) and National Human Genome Research Institute (1R01HG006786).

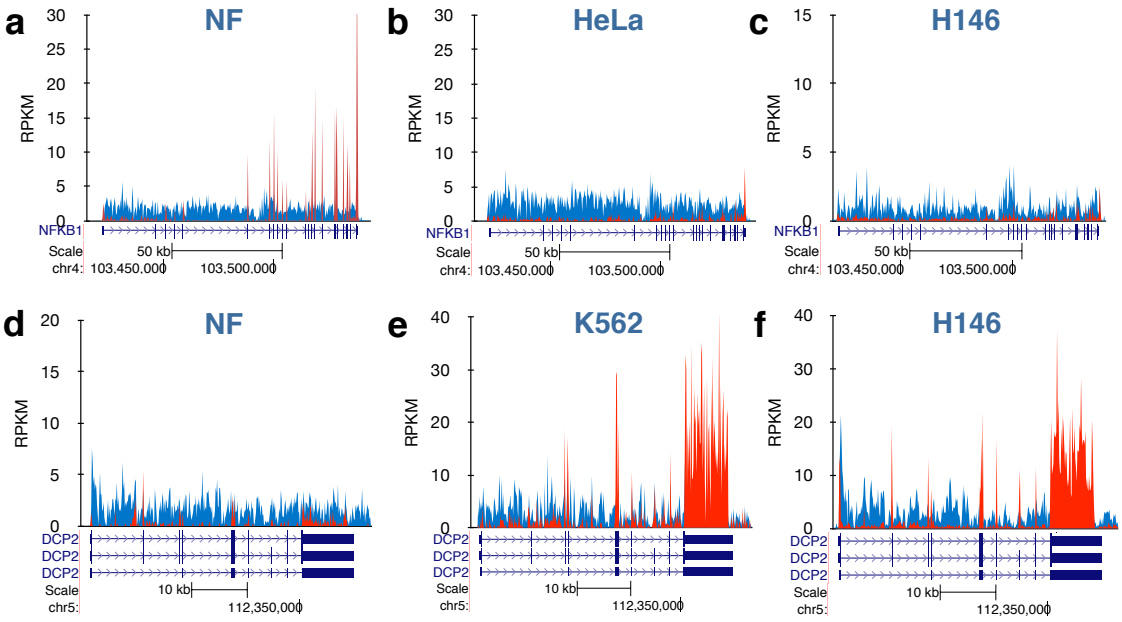


Figure 2.1: BruChase-Seq reveals differential RNA stabilities across cell lines. (a) High exonic reads at 6 h for the NFKB1 transcript in human broblasts indicating high relative stability. Low exonic reads at 6 h of NFKB1 transcripts in (b) HeLa and (c) H146 cells indicating low relative stability. (d) Low exonic reads at 6 h for the DCP2 transcript in human broblasts indicating low relative stability. High level of exonic reads at 6 h of the DCP2 transcript in (e) K562 and (f) H146 cells indicating relative high stability. Nascent RNA corresponds to the blue trace and the 6-h old RNA corresponds to the red trace. The gene maps are from RefSeq Genes hg19 (UCSC genome browser `http://genome.ucsc.edu/`)

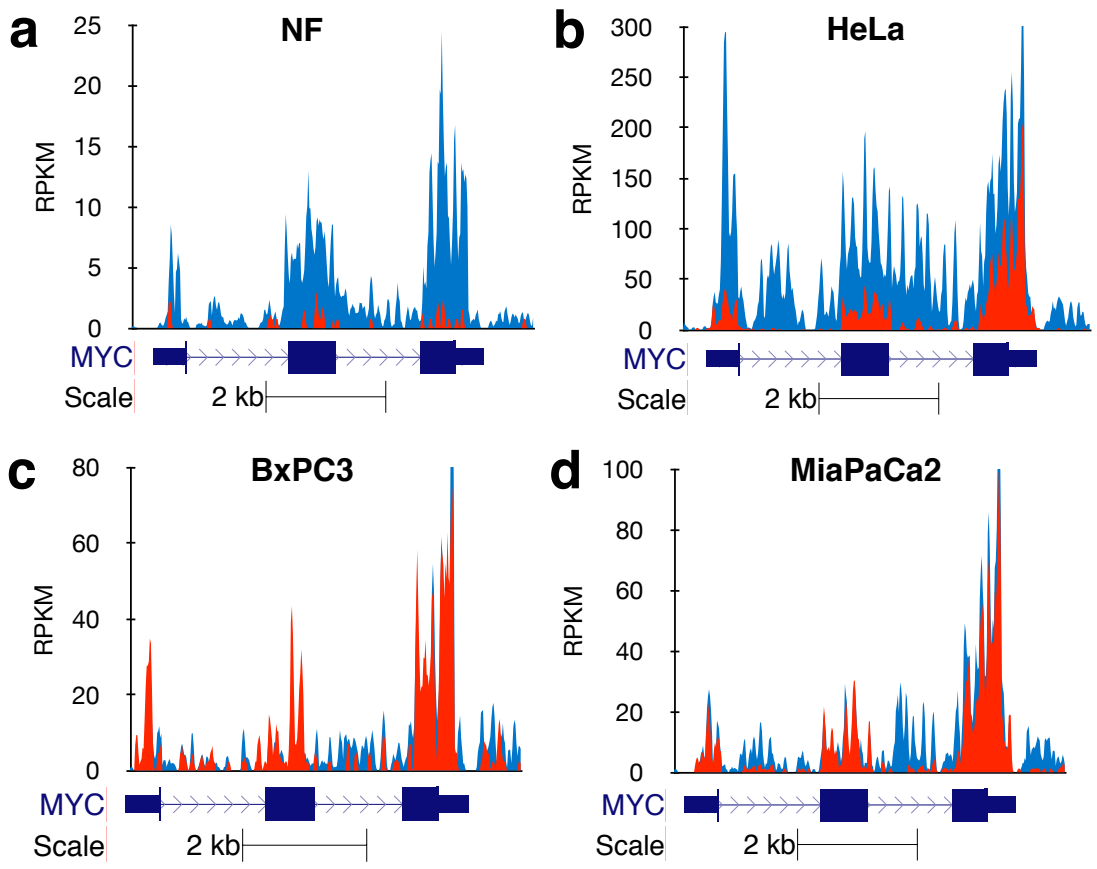


Figure 2.2: The MYC transcript show enhanced stability in the pancreatic cancer lines BxPC3 and MiaPaCa2 as assessed by BruChase-Seq. Low relative stability of MYC transcripts found in (a) human broblasts and (b) HeLa cells assessed by the low level of exonic reads at 6 h in these two cell lines. High relative stability of MYC transcripts in (c) BxPC3 and (d) MiaPaCa2 assessed by BruChase-Seq. Nascent RNA corresponds to the blue trace and the 6-h old RNA corresponds to the red trace. The gene maps are from RefSeq Genes hg19 (UCSC genome browser `http://genome.ucsc.edu/`).

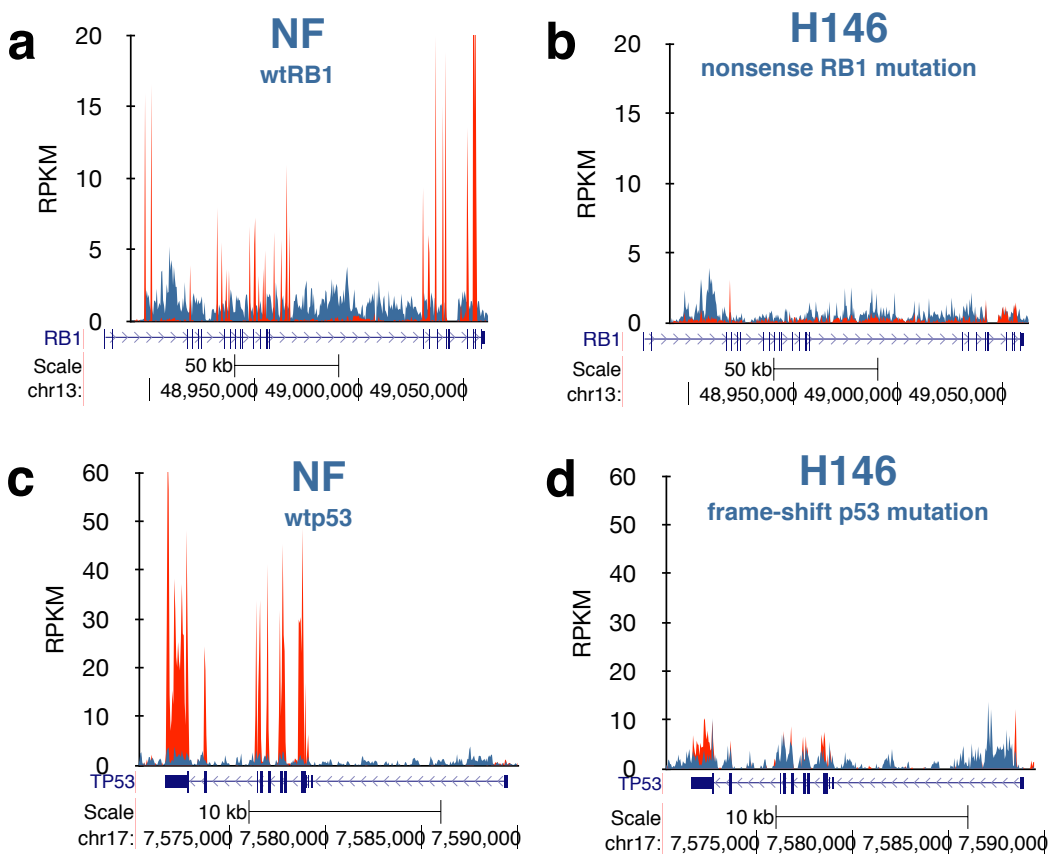


Figure 2.3: Results obtained with BruChase-Seq show reduced stability of mutant transcripts. (a) The wild-type RB1 transcript is stable in human broblasts as assessed by the high exonic reads from the 6-h RNA sample. (b) The RB1 transcript containing a nonsense mutation in H146 cells shows low stability as assessed by the low exonic reads from the 6-h RNA. (c) The wild-type TP53 transcript in human broblasts shows high relative stability while (d) the TP53 transcript with a frameshift mutation in H146 cells shows low stability. Nascent RNA corresponds to the blue trace and the 6-h old RNA corresponds to the red trace. The gene maps are from RefSeq Genes hg19 (UCSC genome browser `http://genome.ucsc.edu/`).

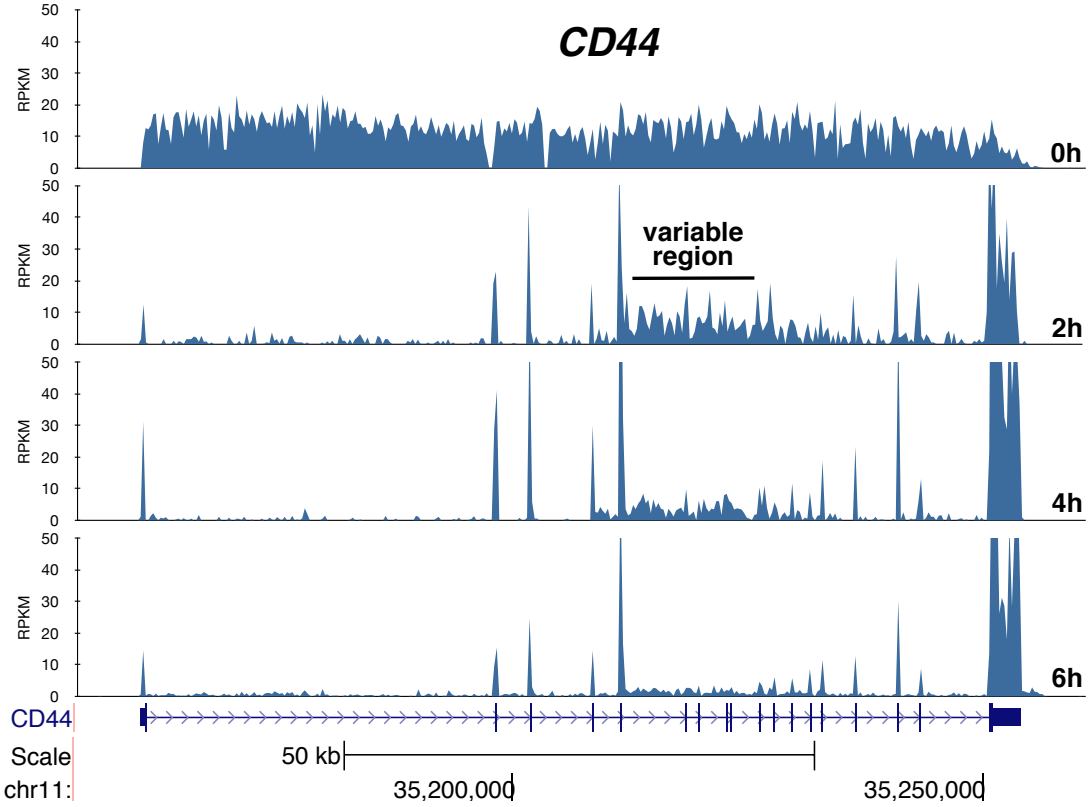
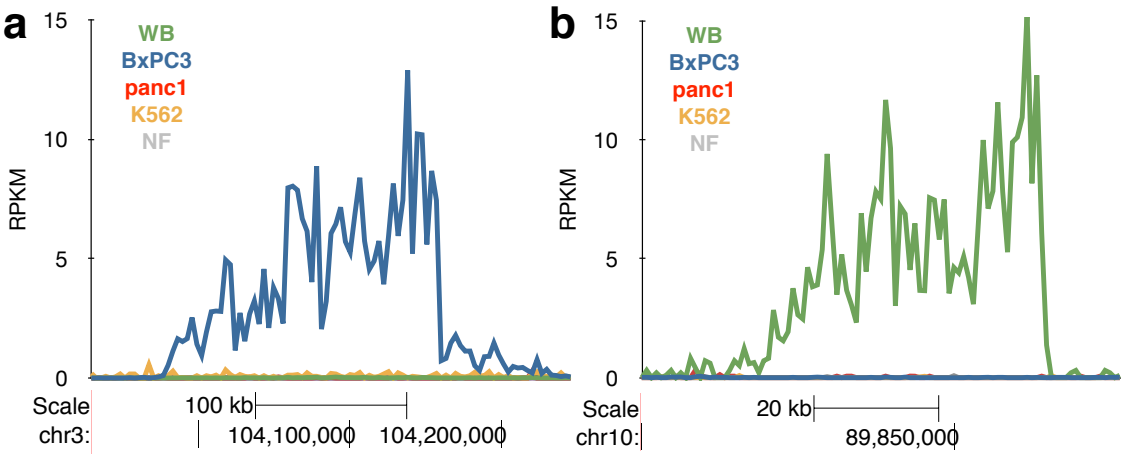


Figure 2.4: Use of BruChase-Seq to assess splicing kinetics of the CD44 transcript. HeLa cells were incubated with 2 mM bromouridine for 30 min to label nascent RNAfollowed by chases in uridine for 0, 2, 4 and 6 h to generate RNA populations of different ages. It can be noted that intronic sequences that are in the region termed "variable region" are removed/degraded much slower than adjacent intronic sequences. The gene maps are from RefSeq Genes hg19 and only one of the many isoforms of the CD44 gene are shown for simplicity (UCSC genome browser `http://genome.ucsc.edu/`).



c

Cell type	Total non-annotated lncRNAs	Cell-specific lncRNAs	shared with WB	shared with BxPC3	shared with panc1	shared with K562	shared with NF
WB	57	41 (72%)	-	2	3	4	2
BxPC3	54	29 (54%)	7	-	10	19	2
panc1	25	9 (36%)	6	11	-	8	10
K562	153	130 (85%)	6	16	10	-	5
NF	21	15 (71%)	1	1	2	1	-

Figure 2.5: Cell type-specic expression of non-annotated lncRNAs identied using Bru-Seq. (a) Example of lncRNA exclusively expressed in BxPC3 cells. (b) Example of lncRNA exclusively expressed in whole blood. (c) Table listing cell types examined, the number of non-annotated lncRNA expressed in the different cell types and the number of lncRNAs shared among the different cell types. For simplicity, the reads in (a) and (b) are shown as positive values with transcription going from right to left.

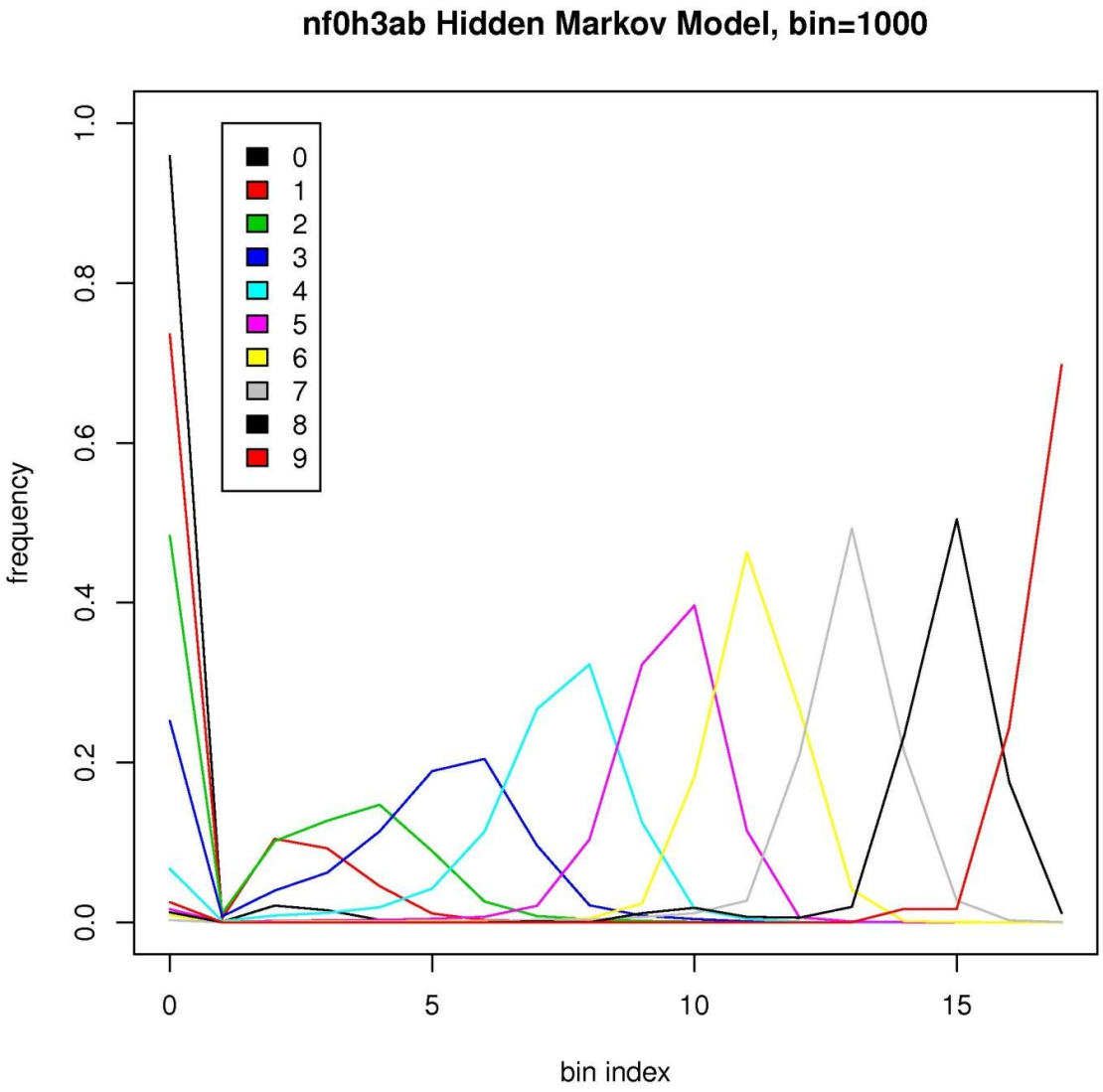


Figure S2.1: Example of the hidden Markov model emission probabilities calculated for sample nf0h3ab, as generated by q master segment. Each trace shows the frequency of different bin input states (indicated on the X axis) for each of ten different output states (indicated in the legend). Probabilities were trained based on observed bin RPKM in annotated genes and subsequently applied genome wide.

CHAPTER III

Coordinated regulation of synthesis and stability of RNA during the acute TNF-induced pro-inflammatory response

3.1 Abstract

Steady-state gene expression is a coordination of synthesis and decay of RNA through epigenetic regulation, transcription factors, miRNAs and RNA-binding proteins. Here we present Bru-Seq and BruChase-Seq to assess genome-wide changes to RNA synthesis and stability in human fibroblasts at homeostasis and after exposure to the pro-inflammatory tumor necrosis factor (TNF). The inflammatory response in human cells involves rapid and dramatic changes in gene expression and the Bru-Seq and BruChase-Seq techniques revealed a coordinated and complex regulation of gene expression both at the transcriptional and post-transcriptional levels. The combinatory analysis of both RNA synthesis and stability using Bru-Seq and BruChase-Seq allows for a much deeper understanding of mechanisms of gene regulation than afforded by the analysis of steady-state total RNA and should be useful in many biological settings.

Official citation:
Paulsen, M.T., Veloso, A., Prasad, J., Bedi, K., Ljungman, E.A., Tsan, Y., Chang, C., Tarrier, B., Washburn, J., Lyons, R., Robinson, D., Kumar-Sinha, C., Wilson, T.E., Ljungman, M. Coordinated regulation of synthesis and stability of RNA during the acute TNF-induced pro-inflammatory response. *PNAS*, 110(6), 2240-5, 2013. doi: 10.1073/pnas.1219192110

3.2 Introduction

The acute inflammatory response is critical for the defense against infections and in the healing of damaged tissues (Medzhitov, 2008). The orchestration of the reprogramming of gene expression associated with the acute inflammatory response is complex and involves both transcriptional and post-transcriptional regulation (Tian et al., 2005; Hao and Baltimore, 2009; Anderson, 2010; Khabar, 2010). Conventional exploration of gene expression using total RNA does not fully capture this complexity since it does not provide insight into the contribution of nascent RNA synthesis or RNA decay to steady-state RNA changes. A number of different approaches have recently been developed to assess nascent RNA synthesis in cells such as GRO-Seq (Core et al., 2008), NET-Seq (Churchman and Weissman, 2011), NUN-RNA-Seq (Khodor et al., 2011), and metabolic labeling of nascent RNA using microarrays (Ohtsu et al., 2008) or RNA-Seq (Rabani et al., 2011; Schwanhusser et al., 2011). By comparing the data obtained with metabolically labeled nascent RNA with the steady-state RNA levels, the rates of degradation of all transcripts can be computationally estimated. The stability of steady-state RNA can also be estimated from the decay rate of steady-state RNA following transcription inhibition (Lam et al., 2001; Raghavan et al., 2002; Gerstein et al., 2012) or by immunoprecipitation of metabolically labeled steady-state RNA following different chase periods (Munchel et al., 2011; Tani et al., 2012). These approaches work well when the system is at homeostasis but not when conditions are altered by environmental stimuli or stress, such as the induction of the acute inflammatory response, when the rates of decay of transcripts are expected to change (Rabani et al., 2011; Schwanhusser et al., 2011).

In this study, we present Bru-Seq and BruChase-Seq based on bromouridine

pulse labeling of nascent RNA followed by chases in uridine to obtain RNA populations of specific ages. The Bru-labeled RNA is then immunocaptured followed by deep sequencing. These new techniques allowed us to assess changes in the rates of both synthesis and degradation of RNA globally following the activation of the pro-inflammatory response by TNF. Our results provide a comprehensive and complex picture of the contribution of transcriptional and post-transcriptional regulation during homeostasis and in the reprogramming of gene expression during the acute TNF-induced pro-inflammatory response.

3.3 Results

3.3.1 Metabolic labeling of nascent RNA with bromouridine

To explore the contributions of both transcriptional and post-transcriptional regulation to the acute pro-inflammatory response, we developed Bru-Seq and BruChase-Seq. These approaches are based on a short labeling (30 min) of nascent RNA with bromouridine (Bru), followed by direct isolation (Bru-Seq) or a chase in uridine for different periods of time (BruChase-Seq). Following labeling and chase, the Bru-containing RNA is specifically isolated from total RNA using anti-BrdU antibodies. This material is then converted into a strand-specific cDNA library (Illumina TruSeq RNA Sample Prep Kit) and subjected to deep sequencing (Illumina Hi-Seq) followed by analysis of mapped read density across the reference genome (Fig. 3.1A). Bromouridine has been used previously to label nascent RNA (Haider et al., 1997; Ohtsu et al., 2008) or steady state RNA (Tani et al., 2012) in cells or in vitro (Core et al., 2008) followed by capturing of Bru-labeled RNA using specific anti-BrdU antibodies and it is less toxic to cells than the analogs 4-thiouridine (4sU) and ethynyluridine (EU) (Tani et al., 2012). The BruChase-Seq approach is unique in that the nascent RNA pool is labeled followed by uridine chases for different periods of time, making

it possible to analyze RNA populations of distinct ages.

Incubation of human fibroblasts with 2 mM bromouridine for 30 min gave a robust RNA incorporation as measured by immunocytochemistry using anti-BrdU antibodies. As expected, this Bru incorporation was reduced by simultaneous incubation with the transcription inhibitors actinomycin D or DRB. Moreover, removal of the bromouridine from the culture plates followed by a chase in 20 mM uridine resulted in the gradual disappearance of the nuclear signal as nascent RNAs are expected to be processed and exported out of the nucleus. The retention of some Bru signal even after a 2-hour chase is most likely due to a continuation of BrUTP incorporation during the beginning of the chase until the supplied uridine in the chase media is converted to UTP and to the extended time required to complete transcription of long genes.

The capturing of Bru-labeled RNA is performed using anti-BrdU antibodies conjugated to magnetic beads. The amount of unlabeled RNA captured as background with these antibody-conjugated beads was estimated to be below 0.4%. Notably, the isolated Bru-labeled RNA had a size distribution that differed markedly from the steady state RNA from which it was captured. This size distribution is similar to what was reported for isolation of nuclear RNA (NUN) in Drosophila cells (Khodor et al., 2011).

3.3.2 Bru-Seq

Sequencing and mapping of cDNA libraries prepared from nascent RNA captured immediately after the 30 min Bru-labeling pulse (0-hour) informs on where in the genome and at what rate transcription occurs across the cell population under study during the labeling period (the transcriptome). As can be seen in Figure 3.1B, the mapped reads from the nascent RNA covered entire genes, including both introns and

exons. This is in sharp contrast to the traditional mapping of steady-state RNA, which contains almost exclusively mature RNAs lacking intronic sequences. For some of the genes shown, reads could be detected on the opposite strand upstream of transcriptional start sites. This signal corresponds well to previously described promoter upstream transcripts (PROMPTs) (Preker et al., 2008), which are typically extremely short lived and not readily detectable using steady-state RNA. This illustrates the power of Bru-Seq for detecting unstable transcripts.

The relative rate of transcription of genes within and between samples can be inferred by integrating the read signal throughout the gene and normalizing for the length of the gene and the total number of reads in the library to obtain "reads per thousand base pairs per one million reads" (RPKM). Due to ongoing splicing and degradation of intronic sequences during the labeling period, the integrated signal across the whole gene will be slightly underestimated. However, when the transcription rate of a particular gene is compared between two samples, this underestimation should not impact the estimation of the fold difference in expression between the two samples. The results obtained with Bru-Seq were highly reproducible when comparing different biological samples of the same cell line, and data from Bru-Seq correlated well with data obtained with the established in vitro run-on technique GRO-Seq on human lung fibroblasts. The transcriptome data for all genes ($> 300bp$) in human fibroblasts using Bru-Seq can be found in table S3. The highest transcribed gene in growing human fibroblasts was MALAT1, which encodes a nuclear ncRNA thought to promote alternative splicing of various transcripts (Tripathi et al., 2010). Over 4,000 annotated genes (20%) were silent at our level of detection in human fibroblasts.

3.3.3 BruChase-Seq

By chasing Bru-treated cells with uridine, RNA populations of defined ages can be obtained and many features of transcription, splicing and RNA degradation can be observed with a clarity not afforded by other techniques. We obtained similar ranking order of the intrinsic stabilities of most transcripts using a 2-hour or a 6-hour chase. However, since the length of a gene, i.e. the time it takes to complete transcription, and the lag time for effective uridine quenching of the BrUTP pool may influence the assessment of RNA stability using short chase periods, we decided to use a 6-hour chase for all subsequent BruChase-Seq experiments. We find close correlation between data obtained with BruChase-Seq and with BruChase-qRT-PCR. As can be seen in Figure 3.1C, the 6h old Bru-labeled SMAD4 RNA was highly enriched for exons, consistent with maturation by splicing. By comparing the RPKM values of the exons in the 6h old sample with the RPKM values throughout the gene in the nascent RNA sample, the intrinsic stability of a transcript can be estimated. An example of a stable transcript is HIF1A (Fig. 3.1D) and an example of an unstable transcript is BTG2 (Fig. 3.1E). A list of relative intrinsic stabilities of transcripts in human fibroblasts can be found in table S3. Finally, highly unstable pre-miRNA transcripts were readily detected when sequencing nascent RNA (Bru-Seq) but not when sequencing 6h old RNA (BruChase-Seq). An example of this is shown in Figure 3.1F for the transcript of the miR23-A, miR24-2 and miR27 cluster that has not been completely annotated previously.

3.3.4 Genome-wide analyses

When analyzing the distribution of sequence reads throughout the genome of human fibroblasts, 10% of the Bru-Seq reads came from exonic regions, 75% from

intronic regions, 3% was antisense RNA and 12% came from unannotated, intergenic regions (Fig. 3.2A). The distribution of reads of the 6-hour old RNA with BruChase-Seq showed that the relative abundance of exonic reads increased to 49% while intronic reads decreased to 24% reflecting the higher stability of exons relative to introns. Using a Hidden Markov Model (HMM) segmentation analysis to map all transcription units independently of prior gene annotations (see Online Materials and Methods), we were able to estimate that about 34% of the fibroblast genome was giving rise to a detectable transcription signal while 66% of the genome did not generate any signal detectable above background (Fig. 3.2B). Thus, at our sequencing depth and growth conditions, we estimate that the "transcriptome" in human fibroblasts is confined to about 34% of the genome which is in concordance with recent reports using a similar HMM segmentation approach for other human cell lines (Djebali et al., 2012).

When plotting the transcriptome against the RNA stabilome of both mRNAs and annotated ncRNAs we did not observe a clear relationship between relative transcription rate and relative RNA stability (Fig. 3.2C). The distribution of transcription rates and stabilities of mRNAs and annotated ncRNAs were fairly similar, suggesting that synthesis and turnover of mRNA and ncRNAs may be regulated by similar mechanisms as recently suggested (Clark et al., 2012; Tani et al., 2012). Performing DAVID gene ontology analysis to test for gene enrichment we found that genes involved in the KEGG pathway "ribosome" were significantly enriched in the highest transcribed gene set (p-value $< 6.84 \times 10^{-54}$) (Fig. 3.2D). Interestingly, the "ribosome" pathway was also highly enriched in the gene set of the least stable transcripts (p-value $< 1.37 \times 10^{-52}$). This finding that transcripts from ribosomal protein genes are very unstable concur with a study in *S. cerevisiae* (Grigull et al., 2004) and sug-

gests a unique mechanism whereby cells regulate ribosome biogenesis. In addition, 14 of the 100 most highly transcribed genes were found to generate transcripts that were among the 100 least stable transcripts. These genes were CYR61, DUSP1, DUSP6, EGR1, EID3, FAM43A, FOS, FOSB, ID1, JUN, JUNB, KLF6, MCL1, and ZFP36.

3.3.5 Analysis of RNA synthesis and stability of mitochondrial and ribosomal RNA

The human mitochondrial genome is circular and consists of 16,569 bp encoding 8 mRNAs, 2 rRNAs and 22 tRNAs (Asin-Cayuela and Gustafsson, 2007). The observation that the steady-state levels of the individual transcripts differ greatly despite being transcribed in a polycistronic fashion indicates that the levels of these transcripts must be under post-transcriptional regulation (Mercer et al., 2011). Using BruChase-Seq we directly assessed the relative stability of the mitochondrial RNAs. While the two mitochondrial-encoded ribosomal RNAs, RNR1 and RNR2, showed high relative stability, the transcripts of the protein-coding genes were highly unstable.

To assess the synthesis and stability of ribosomal rRNA in human fibroblasts, we first collapsed the approximately 400 rDNA genomic repeat sequences into one rDNA sequence and aligned all the reads to this single locus as recently described (Zentner et al., 2011). Significant processing of intergenic spacer RNA in the primary rRNA transcript is apparent. Due to a size selection step of cDNA prior to sequencing, we did not obtain sufficient amounts of the short 5.8S RNA for our analysis. Using BruChase-Seq to assess the stability of the ribosomal RNA transcripts 18S and 28S relative to all other transcripts showed an expected high stability. When calculating the contribution of mitochondrial and ribosomal RNA to the total pool of nascent RNA reads, rRNA made up about 10% and mitochondrial RNA about 7%. These

numbers increased to 38% and 9% respectively when analyzing the pool of 6-hour old RNA reflecting their overall relative stability.

3.3.6 Intron retention

We observed a number of genes that produced transcripts where specific introns were retained even following a 6-hour chase. A strong correlation was found between the intron retention fraction, defined as the signal in an intron relative to the exons of the same gene, and the fraction of reads crossing the boundaries of the intron that were unspliced, indicating that these intronic reads originated from introns retained in the transcripts rather than from reduced rates of degradation of spliced introns. We found 360 introns that were retained to more than 10% in the 6-hour old RNA. Of these introns, 116 were found on genes with at least one additional retained intron suggesting that when one intron is poorly spliced there is a high likelihood that an additional intron will be poorly spliced. According to DAVID gene ontology analysis, the gene list of transcripts with retained introns was highly enriched in "phosphoproteins" (p-value $< 1.7 \times 10^{-20}$). We are currently analyzing splicing kinetics genome-wide using multiple chase time points.

3.3.7 The TNF-induced transcriptome

We next applied Bru-Seq to explore alterations in the transcriptome following an acute exposure to the inflammatory cytokine TNF in human fibroblasts. It is well known that the pro-inflammatory response involves dramatic changes in RNA levels in cells and these changes are thought to be due both to transcriptional and post-transcriptional regulation (Tian et al., 2005; Hao and Baltimore, 2009; Anderson, 2010; Khabar, 2010). We first performed time course experiments with Bru-labeling of nascent RNA and observed dramatic induction of nascent RNA synthesis already

after 30 minutes of TNF-treatment. The induction of nascent RNA synthesis for these genes peaked at 2 to 6 hours after addition of TNF (Fig. 3.3 A&B). Performing a genome-wide Bru-Seq data analysis of TNF-induced and repressed genes using DESeq analysis (Anders and Huber, 2010) we found that 472 genes were up regulated and 204 genes down regulated at least 2-fold following a one hour incubation with TNF. Examples of up regulated genes were IL1A and IL1B while HES1 and KLF4 represent genes down regulated at the level of RNA synthesis (Fig. 3.3C-F).

3.3.8 The TNF-induced RNA stabilome

The rapid changes in gene expression following the induction of the pro-inflammatory response in human fibroblasts treated with TNF have been shown to depend on the induction of synthesis of genes with low intrinsic transcript stability, (Hao and Baltimore, 2009). We first assessed the intrinsic stability of inflammation-associated transcripts using the bromouridine pulse-chase strategy coupled to real-time RT-PCR arrays and we confirmed that many of the pro-inflammatory genes generated very unstable transcripts (Fig. 3.3G) (Hao and Baltimore, 2009). We next assessed whether exposure to TNF may affect the stability of these transcripts in unperturbed cells using the bromouridine pulse-chase approach. Cells were pulse-labeled for 30 min in the absence of TNF and were then chased for 6 hours in the presence or absence of TNF and as can be seen, TNF dramatically increased the stability of many of these transcripts (Fig. 3.3H). We next used BruChase-Seq to examine the effect of TNF on RNA stability genome-wide and detected significantly increased stabilities of 152 transcripts, such as SOD2 and ICAM1 (Fig. 3.3I&J). We also observed 58 transcripts significantly destabilized by TNF treatment after Bru-labeling, such as GAS1 and HOXA9 (Fig. 3.3K&L). Other members of the HOXA gene cluster, such as HOXA6, HOXA11, HOXA13 and HOTAIR, also showed reduced transcript

stability following TNF-treatment during the chase.

3.3.9 Coordinated and complex regulation of the transcriptome and RNA stabilome after TNF

The results show that cells induce the acute pro-inflammatory response by regulating both synthesis and stability of RNA. Some genes were found to be up regulated transcriptionally, post-transcriptionally or both. Other genes were down regulated transcriptionally, post-transcriptionally, or both, or through a mixture of up and down regulation. We also observed dramatic induction of primary transcripts of miR155, miR146A and miR3142 and repression of primary transcripts of miR143, miR145 and miR614 following a 1-hour TNF treatment. MIR155 and MIR146 have been shown to be induced by NFkB during inflammation (Taganov et al., 2006; O'Connell et al., 2007) and MIR155 has been shown to suppress MIR143 (Jiang et al., 2012).

Finally, in very large genes with affected transcription following TNF treatment we could "visualize" the wave of induced or repressed transcription moving through the gene. For the 300 kb long FNDC3B we observed that the front of the induced transcription wave had reached about 260 kb into the gene during the 90 minute experiment (Fig. S3.1). For the TOX gene we saw reduced RNA synthesis in the first 180 kb into the gene suggesting reduced initiation and then the spread of reduced transcription into the gene. The transcription signal from the SAMD4A gene showed a "hump" suggesting that this gene was transiently induced following TNF exposure. This illustrates the power of Bru-Seq in capturing dynamic events such as initiation and elongation of transcription.

Performing DAVID gene ontology enrichment analysis on the genes that were induced at the transcriptional and/or post-transcriptional level by TNF treatment

we found a number of pathways affected by TNF (Fig 3.4). Bru-Seq showed over 470 genes induced by TNF and they were enriched in pathways that are known to be induced as part of the acute pro-inflammatory response such as "inflammation", "cytokine production" and "anti-apoptosis". In addition, Bru-Seq detected over 200 genes being repressed rapidly after TNF exposure and these genes were enriched in pathways such as "negative regulation of transcription", "nucleosome core" and "ubiquitin conjugation". Using BruChase-Seq we found that over 200 genes were regulated post-transcriptionally following TNF treatment and some of these pathways were in common with those affected transcriptionally such as "inflammatory response", "response to wounding" and "antiapoptosis".

3.4 Discussion

TNF is an important pro-inflammatory cytokine that mediates its biological effects by activating NFkB, AP-1 and p38 (Beg and Baltimore, 1996; Aggarwal, 2003). NFkB and AP-1 are transcription factors regulating transcription initiation while p38 is a kinase that has been shown to regulate mRNA stability by phosphorylating RNA-binding proteins such as tristetraproline (TTP) (Anderson, 2010). Numerous studies have profiled TNF-induced gene expression and mRNA stabilization in different cells using steady-state RNA and the transcription inhibitor actinomycin D. However, actinomycin D induces cellular stress responses involving p53 (Ljungman et al., 1999) and have been shown to introduce artifacts in mRNA stability determinations Dölken et al. (2008); Munchel et al. (2011). In this study we developed Bru-Seq and BruChase-Seq to profile the transcriptome and RNA stabilome of unperturbed human skin fibroblasts at homeostasis and following induction of the pro-inflammatory response by TNF treatment. While the two new techniques confirmed

changes to genes known to respond to TNF, which thus validated the techniques, the novelty of this study lies in the comprehensive nature of the global analysis of both synthesis and stability of RNA during the acute pro-inflammatory response.

Our results revealed that the TNF-induced pro-inflammatory response elicits a coordinated and complex reprogramming of gene expression by induction or repression of transcription and/or RNA stability. It was noticeable that many of the pro-inflammatory cytokines and chemokines were induced both transcriptionally and post-transcriptionally by TNF. It is possible that this dual induction occurs as a result of the activation of two separate signaling arms, such as NFkB for induction of gene-specific transcription and p38 kinase for promoting RNA stabilization via phosphorylation of specific RNA-binding proteins. Alternatively, the reduced decay of these transcripts may be due to "mass action" where the machinery that normally targets these transcripts for degradation becomes overwhelmed by the dramatically increased amounts of transcripts generated. Future studies will be aimed at distinguishing between these models.

The data obtained with the Bru-Seq and BruChase-Seq techniques provides a record of both ongoing transcription (transcriptome) and the rate of decay of the generated RNA (RNA stabilome). In addition, the approaches can determine splicing efficiencies genome-wide and detect and map the generation of short-lived RNA species such as PROMTs and pre-miRNAs. Since the Bru-Seq approach only measure newly made RNA, rapid reduction in transcription rates can be estimated without relying on the decay of pre-existing RNAs. Thus, our list of genes rapidly inhibited following TNF treatment is novel and should contribute to the understanding of the acute pro-inflammatory response. We believe that the Bru-Seq and BruChase-Seq techniques should have a wide utility in many biological settings were transcriptional

and post-transcriptional regulation is desired to be assessed on a genome-wide scale.

3.5 Material and Methods

3.5.1 Bromouridine pulse-chase labeling and isolation of Bru-RNA

Bromouridine (Aldrich) was added to the media of normal diploid fibroblasts to a final concentration of 2 mM and cells were incubated at 37C for 30 min. Cells were then washed 3 times in PBS and either collected directly (nascent RNA, Bru-Seq) or chased in conditioned media containing 20 mM uridine for 6 hours at 37C (6-hour old RNA, BruChase-Seq). For TNF treatments, recombinant human TNF-alpha (R&D Systems, Minneapolis, MN) was added to a concentration of 10 ng/ml from a 10 g/ml stock solution in PBS either one hour before (and included during) Bru-labeling (Bru-Seq) or directly following Bru-labeling during the 6h uridine chase (BruChase-Seq). Total RNA was isolated using TRIzol reagent (Invitrogen) and Bru-labeled RNA was isolated from the total RNA by incubation with anti-BrdU antibodies (BD Biosciences) conjugated to magnetic beads (Dynabeads, Goat anti-Mouse IgG, Invitrogen) under gentle agitation at room temperature for 1 hour. For more detail, please see Online Materials and Methods.

3.5.2 cDNA library preparation and Illumina sequencing

Isolated Bru-labeled RNA was used to prepare strand-specific DNA libraries using the Illumina TruSeq Kit (Illumina) according to the manufacturers instructions with modifications noted in Online Materials and Methods. Sequencing of the cDNA libraries prepared from nascent RNA or 6-hour old RNA was performed at the University of Michigan Sequencing Core using the Illumina HiSeq 2000 sequencer.

3.5.3 Data analysis

Base calling was performed by the University of Michigan DNA Sequencing Core using Illumina Casava v1.8.2. and read mapping was performed using TopHat accepting only reads that could be mapped uniquely to the genome. For determining exon, intron, and gene coverage, a single condensed transcriptome map of the genome was constructed. Bedtools was then used to determine the coverage within each exon and intron region, similar to base and bin coverage. Subsequently, coverage values for all exons of each gene were summed, as well as separately all introns of each gene. Finally, values for all exons and all introns were summed to obtain the coverage for the entire gene. For comparing and ranking genome feature coverage, we calculated RPKM values as previously described (Mortazavi et al., 2008). For identifying transcribed genome regions independently of any prior annotation, we performed genome segmentation using a Hidden Markov model (HMM). For more details, please see Online Materials and Methods.

3.5.4 Data availability

The primary data used in the analyses will be deposited at NCBI's Gene Expression Omnibus.

3.6 Acknowledgements

We thank Manhong Dai and Fan Meng for administration and maintenance of the University of Michigan Molecular and Behavioral Neuroscience Institute (MBNI) computing cluster and the personnel at the University of Michigan Sequencing Core for technical assistance. This work was supported by funds from University of Michigan Bioinformatics Program, University of Michigan Biomedical Research Council, the Will and Jeanne Caldwell Endowed Research Fund of the University of Michi-

gan Comprehensive Cancer Center, University of Michigan School of Public Health (NIEHS P30), Department of Defense, Uniting Against Lung Cancer, University of Michigan Nathan Shock Center, University of Michigan Office of the Vice President of Research, National Cancer Institute (5R21CA150100), National Institute of Environmental Sciences (1R21ES020946) and National Human Genome Research Institute (1R01HG006786).

3.7 Online Methods

3.7.1 Cell lines, TNF treatment and bromouridine pulse-chase labeling

Diploid human foreskin fibroblasts (gift from Dr. Mary Davis, Department of Radiation Oncology, University of Michigan) were hTERT immortalized and grown as monolayers in MEM supplied with 10% fetal bovine serum and antibiotics (Invitrogen). Bromouridine (Aldrich) was added to the media to a final concentration of 2 mM and cells were incubated at 37°C for 30 min. Cells were then washed 3 times in PBS and either collected directly (nascent RNA, Bru-Seq) or chased in conditioned media containing 20 mM uridine for 6 hours at 37°C (6-hour old RNA, BruChase-Seq). Incubation of human fibroblasts with 2 mM bromouridine for 30 min gave a robust incorporation as measured by immunocytochemistry using anti-BrdU antibodies. As expected, this Bru incorporation was blocked by simultaneous incubation with the transcription inhibitors actinomycin D or DRB. Moreover, removal of the bromouridine from the culture plates followed by a chase in 20 mM uridine resulted in the gradual disappearance of nuclear signal as nascent RNAs were processed and exported out of the nucleus. The retention of some Bru signal even after a 2-hour chase is most likely due to the extended time needed for long genes to complete transcription. For TNF treatments, recombinant human TNF-alpha (R&D Systems, Minneapolis, MN) was used in a concentration of 10 ng/ml from a 10 $\mu g/ml$ stock

solution in PBS.

3.7.2 Isolation of total RNA using TRIzol reagent

Following the bromouridine labeling, the culture medium was aspirated and cells were rinsed with PBS, trypsinized and collected in 10 ml ice-cold media and put on ice. The cells were then counted and equal numbers of cells from each sample (at least 2 million cells) were spun down. The medium was aspirated and the cell pellets were resuspended in 1 ml PBS and spun again. The PBS was aspirated and the cell pellet resuspended in 3 ml of TRIzol reagent (Invitrogen). The samples were homogenized by vigorous pipetting and stored at −80°C until further processing.

To isolate total RNA, the frozen samples were thawed and vortexed for 30 s. To the 3 ml TRIzol lysates, 0.6 ml chloroform was added and the samples were shaken vigorously for 10 s. The caps were removed and the tubes were sealed with parafilm and centrifuged at 12,000 g for 15 min at 4°C. The upper clear aqueous phase, containing RNA, was transferred to new tubes and 1.5 ml of isopropanol was added to each tube and the samples gently agitated. The tubes were sealed with parafilm and centrifuged at 12,000 g for 10 min at 4°C. The supernatants were aspirated and 3 ml of ice-cold 75% ethanol added to each tube and centrifuged at 7,500 g for 5 min at 4°C. The supernatants were carefully aspirated immediately and the tubes placed upside down for RNA pellets to dry. The dry pellets were then dissolved in 200 μl DEPC-treated water and the samples were heated at 55 for 10 min to fully dissolve the RNA. From each sample, 20 μl aliquots were put aside for RNA quantification (NanoDrop, Thermo Scientific) and for potential future analysis of steady-state RNA levels. The total RNA samples were stored at −80°C or directly used for Bru-labeled RNA isolation.

3.7.3 Conjugation of anti-BrdU antibodies to magnetic beads

Magnetic beads (Dynabeads, Goat anti-Mouse IgG, Invitrogen) in storage buffer were transferred to Eppendorf tubes (50 μl bead slurry for each sample to be processed). The magnetic beads were captured using a magnetic stand and the storage buffer was aspirated. The captured beads were then washed 3 times with 0.1% BSA in PBS and resuspended in 0.1% BSA in PBS. To each sample (50 μl of original bead slurry), 2 μg (4μl) of anti-BrdU monoclonal antibodies (BD Biosciences) and 1 μl RNase inhibitor (Invitrogen) were added. The magnetic beads and the antibodies were gently mixed for 1 hr at room temperature followed by 3 washes with PBS containing 0.1% BSA. The beads were then resuspended in PBS containing 0.1% BSA and RNase inhibitor and the beads were distributed evenly to new tubes (200 μl/sample).

3.7.4 Isolation of Bru-containing RNA

The samples containing the isolated total RNA were heated for 10 min at 80°C to denature any double-stranded RNA structures. The heated RNA samples were then added to the tubes with the prepared antibody-conjugated magnetic beads and incubated with gentle rotation for 1 hr at room temperature. The beads were washed 3 times with 200μl PBS containing 0.1% BSA, rotating for 5 minutes for each wash. The captured beads were resuspended in 50μl of DEPC-treated water and transferred to new tubes to avoid any non-specific RNA adhering to the walls of the old tubes. The samples were boiled for 10 min to release the Bru-labeled RNA and the samples cooled and briefly centrifuged before the beads were magnetically captured. The supernatants were transferred into new tubes and stored at −80°C until further use. The fraction of Bru-labeled RNA isolated in this procedure makes up 2-10% of total

RNA depending on the cell type and chase time used.

The amount of unlabeled RNA captured as background with these antibody-conjugated beads was estimated to be below 0.4% (table S1). Notably, the isolated Bru- labeled RNA had a size distribution that differed markedly from the steady state RNA from which it was captured. Very little of the immunoprecipitated RNA was longer than the 28S rRNA (5 kb) while the average length was similar to that of the 18S RNA (1.9 kb). This size distribution is similar to what was reported for isolation of nuclear RNA (NUN) in Drosophila cells (Khodor et al., 2011).

3.7.5 cDNA library preparation

Bru-labeled RNA was mixed with first strand buffer and random primers and fragmented by heating at 85°C for 10 minutes. The first strand cDNA was then synthesized, in the presence of Actinomycin D to result in strand specific reads when indicated in table S2. After purifying the first strand cDNA using AMPure RNAclean beads (Beckman Coulter), the second strand cDNA was synthesized. The resulting cDNA was purified with AMPure XP beads, after which the Illumina TruSeq RNA Sample Prep Kit was used to repair the cDNA ends, adenylate and ligate adaptors to the cDNA. The samples were then run on a 3% agarose gel and size-selected by excising gel slices in the 300bp region. These gel slices were purified using the QIAEX II Gel Extraction Kit (Qiagen) and then the Illumina TruSeq Kit PCR reagents were used to enrich the DNA fragments. After a final purification using AMPure XP beads, the quality and concentration of the DNA libraries were determined using an Agilent Bioanalyzer.

3.7.6 Illumina Hi-Seq sequencing

Sequencing of the cDNA libraries prepared from nascent RNA or 6-hour old RNA was performed at the University of Michigan Sequencing Core using the Illumina HiSeq 2000 sequencer according to manufacturer guidelines.

3.7.7 Read mapping

Base calling was performed by the University of Michigan DNA Sequencing Core using Illumina Casava v1.8.2. All steps after base calling were performed on the Linux cluster maintained by the University of Michigan Molecular and Behavioral Neuroscience Institute. In addition to standard Linux command line utilities, the following software was used in the Bru-Seq and BruChase-Seq data analysis pipeline: Tophat v1.3.2, Samtools v0.1.18. Bedtools v2.15.0, create_transcriptome_map.pl v1.0.0, extractKmers.pl v1.0.0, smooth.pl v1.0.0, and segment.pl v1.0.0, where the latter four Perl scripts written for this work are available via:

`http://tewlab.path.med.umich.edu/software/utilities/utilities.html`.

Throughout, the reference genome was hg19/Build 37, with the NCBI RefGene annotation of transcripts and isoforms serving as a guide to known genes.

Read mapping was performed using TopHat (Trapnell et al., 2009), accepting only reads that could be mapped uniquely to the genome. Reads were allowed to split between annotated exons, even for BruSeq 0-hour nascent RNA samples, but de novo splice junction calling was not performed. A coverage determination was then made for every base in the genome using Bedtools (Quinlan and Hall, 2010) such that a base covered by one read was recorded as having a coverage of 1/read_length. Values for all reads covering a base were summed to obtain the total coverage, noting that all coverage values here and below can be fractional. The genome was then grouped

into bins by summing the base coverage over contiguous genome spans, with the bin of a base set as round(position/bin_size)*bin_size. Strand reversal, to account for the fact that the library construction caused the 1st strand cDNA to be sequenced, was applied to all coverage files, but Tophat BAM files contain reads as initially mapped. A description of all of the samples used and their mapping data can be found in table S2.

3.7.8 Gene synthesis and stability

For determining exon, intron, and gene coverage, a single condensed transcriptome map of the genome was constructed (create_transcriptome_map.pl) by merging all annotated transcripts such that, whenever possible, each genome base on each strand was assigned to one and only one gene region of type intergenic, exon_sense, intron_sense, exon_antisense, intron_antisense, or ambiguous. When conflicts arose between genes, annotation preference was given to a gene on the sense strand. Thus, genes on opposite strands were not considered to be overlapping. When conflicts arose between genes, or gene isoforms, on the same strand, preference was given to annotating the base as an exon, so that no read that might be exonic would be counted as intronic, which is important when examining BruChase-Seq data. When conflicts could not be resolved (e.g. a region contained within an intron of two different genes on the same strand) the base was called ambiguous. Ambiguous regions were omitted from all subsequent gene coverage determinations so that values included only bases identifiable as belonging to a specific gene. Bedtools (Quinlan and Hall, 2010) was then used to determine the coverage within each exon and intron region, similar to base and bin coverage. Subsequently, coverage values for all exons of each gene were summed, as well as separately all introns of each gene. Finally, values for all exons and all introns were summed to obtain the coverage for the entire

gene.

For comparing and ranking genome feature coverage, we calculated RPKM values as described (Mortazavi et al., 2008). RPKM values from Bru-Seq data were taken to reveal the relative level of nascent RNA labeling, and thus ongoing transcription, i.e. synthesis, for different genome features or feature sets. Relative transcript stability is expressed as the ratio of BruChase-Seq/6-hour (exons only) to Bru-Seq/0-hour (exons plus introns) RPKM values to account for the fact that signal levels at 6-hour post-labeling are dependent on stability as well as the very different levels of gene synthesis during the labeling period. Gene coverage for unspliced/0-hour samples was calculated using exons plus introns since both are present in nascent RNAs and introns in fact typically account for the majority of a gene's signal. In contrast, gene coverage for spliced/6-hour samples was determined using only exons since introns have largely been spliced away by 6-hour post-labeling and therefore generate little or no relevant signal. Importantly, RPKM values (synthesis) or RPKM ratios (stability) serve only to rank genes; they do not represent rates or half-lives. In table S3, synthesis is shown for genes $>$ 300 bp, and, for stability, where the 0-hour RPKM value exceeded 0.5 to ensure reliable numerical assessments.

3.7.9 Genome segmentation into transcription units

For identifying transcribed genome regions independently of any prior annotation, we performed genome segmentation using a Hidden Markov model (HMM). Bin coverage values at a bin size of 1000 bp were first subjected to minimal smoothing (smooth.pl, J=1) using the wavelet algorithm as described (Day et al., 2007). Prior to smoothing, bins were identified for which the fraction of mappable bases was less than 10%, where extractKmers.pl was used to parse the genome into all potential reads for determining uniqueness at each genome position, given the requirement

that reads map uniquely to the genome. Such bins were disregarded during segmentation and the remaining bins were adjusted so that the corrected coverage = coverage/mappable fraction. Corrected bin data were then indexed into a series of bounded integer observation values from 0 through 17, logarithmically spaced across bin RPKM values from < 0.0005 to > 100 (table S4). Bins with very low hit densities, including many bins with no coverage, accumulated in index 0 while rare bins with exceptionally high hit densities accumulated in index 17. Emission probabilities were trained from sample data and annotated genes, where the expression level of a gene was used to infer its state index similarly to bin observations but using a smaller set of 10 indices. A frequency distribution of the observation indices for all bins in genes of a given state was used as the estimate of the emission probabilities. Bin to bin transition probabilities were established by a single persistence parameter of 0.995, defined as the probability of remaining in state; all inter-state transitions were equally weighted. The starting probability of state 0, i.e. no transcription, was estimated at 0.5; all other states were equally weighted. The most likely sequence of bin states for the resulting HMM and the set of bin observations was finally determined using the Viterbi algorithm (segment.pl), which established genome segments of inferred transcription states (table S4). Using this segmentation approach we defined over 38,000 segments generating reads above background throughout the genome of human fibroblasts (tables S4). These segments do not all represent unique transcription units since some genes generate, in addition to a segment over the body of the gene, a less intense segment covering a region of transcriptional read-through past the 3'-end termination site.

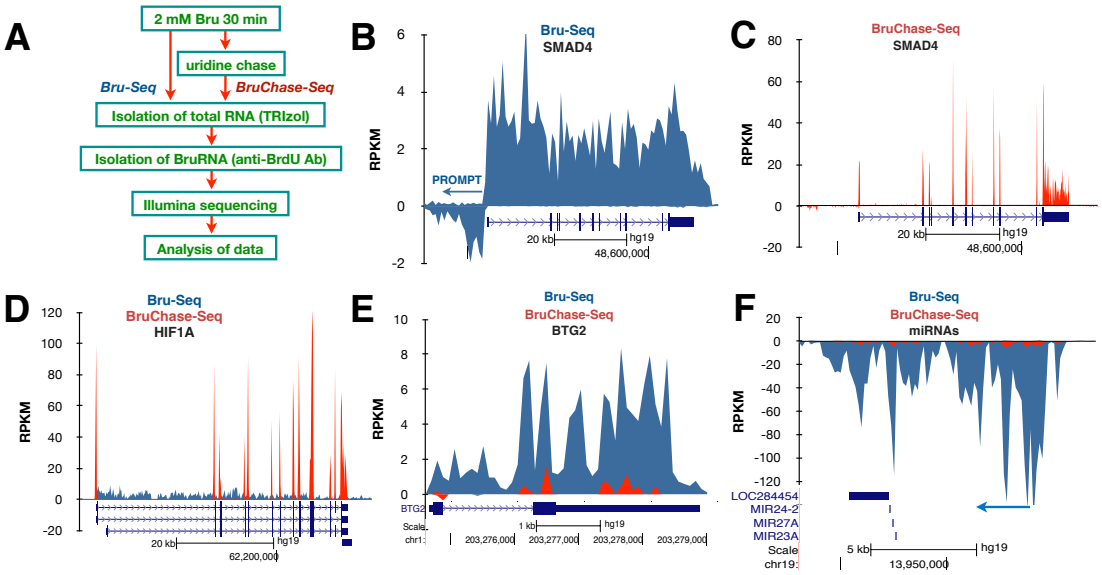


Figure 3.1: Comparisons of nascent and 6-h-old RNA from human fibroblasts using Bru-Seq and BruChase-Seq. (A) Diagram illustrating the main steps in Bru-Seq and BruChase-Seq (see text for details). (B) Sequencing reads from nascent RNA (Bru-Seq) mapping to the SMAD4 gene with reference sequence annotation below with exons and UTRs denoted as black lines. It can be noted that the nascent RNA maps to intronic and exonic sequences and to sequences beyond the 3-end of the gene. Also, mapping of sequence reads on the opposite strand upstream of the SMAD4 transcription start site represents divergent PROMPTs. (C) Sequence reads from 6-h-old SMAD4 RNA (BruChase-Seq) (D) The ratio of the exonic signal in the 6-h old RNA to the signal throughout the gene in the nascent RNA reflects the relative stability of the mature RNA. The mature HIF1A transcript is an example of a stable transcript, whereas the BTG2 transcript (E) is an example of an unstable transcript. (F) The primary transcripts of the miR24-2, miR27A, and miR23A microRNAs are clearly captured by using Bru-Seq but not when analyzing the 6-h-old RNA with BruChase-Seq implicating that the primary miRNA transcripts are rapidly processed into mature miRNAs and size excluded from our analysis. The gene maps are from RefSeq Genes (University of California, Santa Cruz (UCSC) genome browser, `http://genome.ucsc.edu/`)
.

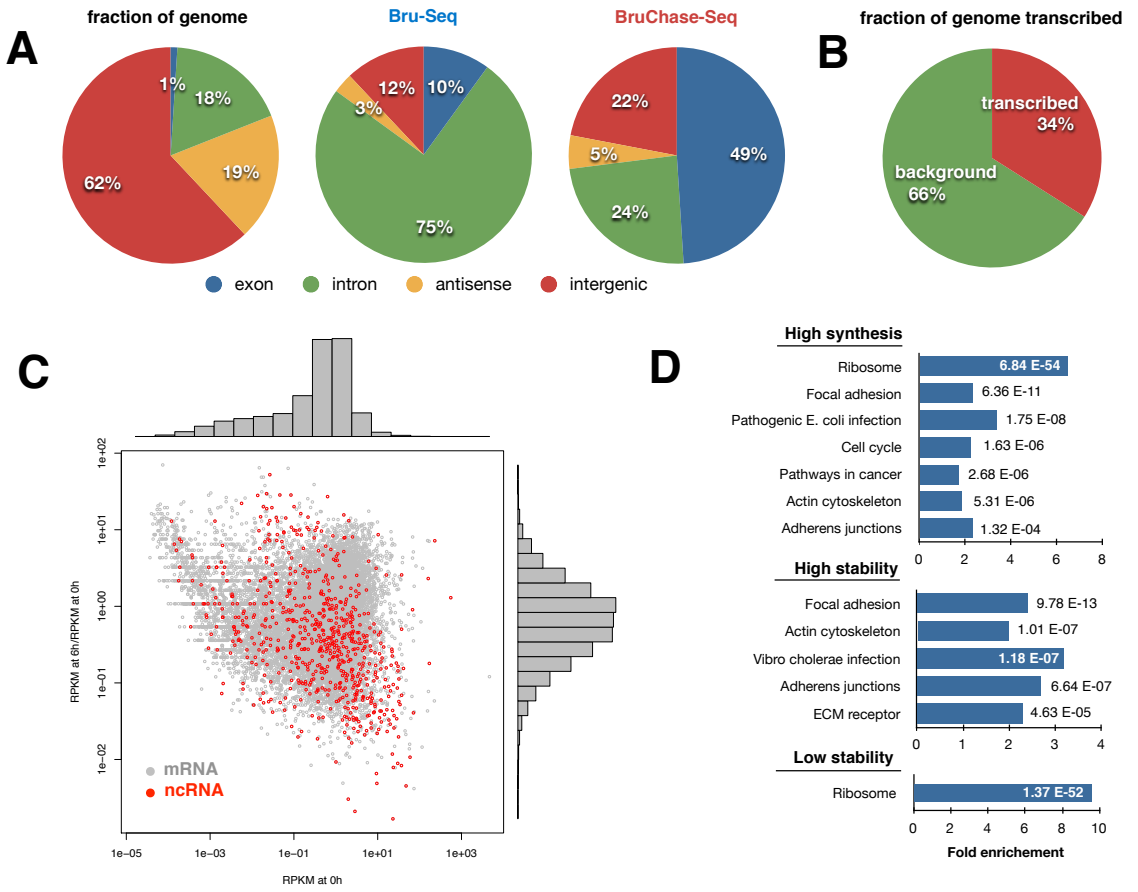


Figure 3.2: Genomic distribution of sequencing reads obtained with Bru-Seq and BruChase-Seq. (A Left) The relative size of exonic, intronic, antisense, and intergenic compartments in the human genome. (A Center) Relative distribution of sequencing reads in these four compartments for nascent RNA (Bru-Seq). (A Right) Relative distribution of sequencing reads from the four different compartments for 6-h-old RNA (BruChase-Seq). (B) Assessment of the portion of the genome generating transcripts using a HMM segmentation analysis. (C) The transcriptome vs. the RNA stabilome with the RPKM values for synthesis (0-h) plotted against the relative stability score (6-h/0-h) for mRNAs (gray circles) and ncRNAs (red circles). (D) KEGG pathway gene enrichment analysis using the DAVID bioinformatics resource showing fold enrichment and P values for pathways enriched in the top 2,000 highly transcribed genes (Top), the top 2,000 most stable transcripts (Middle), and the 2,000 least stable transcripts (Bottom).

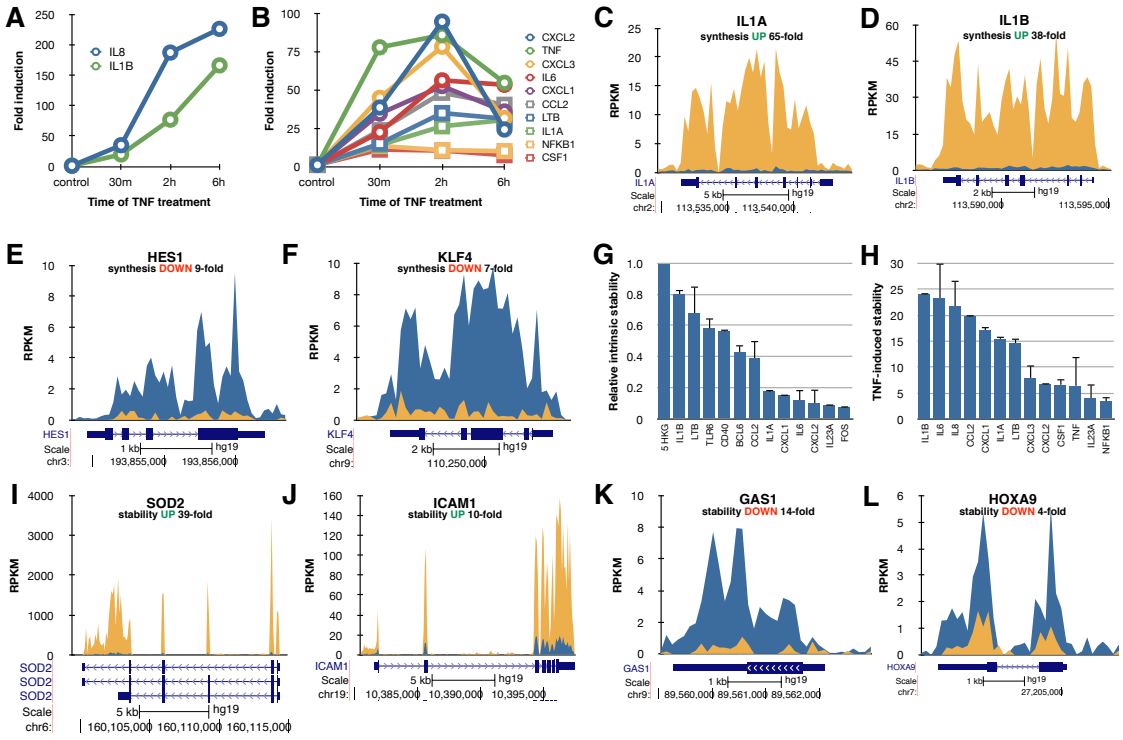


Figure 3.3: Effects of TNF on the synthesis and stability of RNA. (A and B) Human fibroblasts were treated with 10 ng/mL TNF for different periods of time at 37°C with 2 mM bromouridine present during the last 30 min. Total RNA was isolated and Bru-containing RNA isolated by using anti-BrdU antibodies and analyzed by using real-time RT-PCR array technology (inflammation and autoimmunity RT-PCR array; SABiosciences). The values represent the average of two independent experiments. (C and D) TNF-induced transcription of IL1A and IL1B. Blue color represent control, and yellow represent a 60+30 min treatment with TNF. (E and F) Rapid down-regulated transcription of the HES1 and KLF4 genes by TNF. (G) Intrinsic RNA stability of inflammatory cytokine RNAs. Human fibroblast were incubated with 2 mM bromouridine for 30 min followed by a 6-h uridine chase, isolation of Bru-containing RNA from total RNA and real-time PCR analysis using RT-PCR array technology (inflammation and autoimmunity RT-PCR array; SABiosciences). The values are normalized to five housekeeping genes (5 HKG) on the array, which are set to 1.00, and they represents the average of two independent experiments with error bars showing the SD. (H) Same as in G but 10 ng/mL TNF was added at the beginning of the 6-h chase period. The relative abundance of a particular RNA after the 6-h chase in the presence of TNF was compared with its relative abundance after a 6-h chase in the absence of TNF. (I and J) TNF increased stability of the SOD2 and ICAM transcripts. Blue color represents control 6-h chase and yellow represent TNF treatment during the 6-h chase (K and L) TNF treatment resulted in the de-stabilization of the GAS1 and HOXA9 transcripts. The gene maps are from RefSeq Genes (UCSC genome browser, `http://genome.ucsc.edu/`).

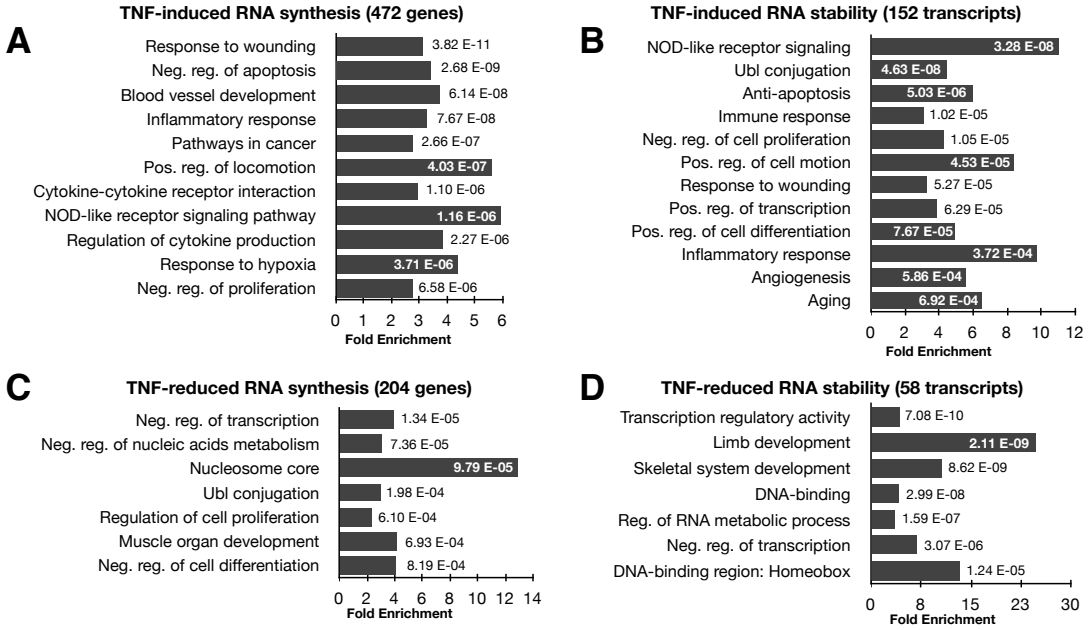


Figure 3.4: Pathway enrichment analysis using DAVID gene ontology for genes affected transcriptionally or posttranscriptionally at least twofold by TNF treatment. (A) Pathway enrichment of genes induced transcriptionally (472 genes) or (B) posttranscriptionally (152 transcripts). (C) Pathway enrichment for genes repressed transcriptionally (204 genes) or (D) posttranscriptionally (58 transcripts). The bars represent fold enrichment of the particular pathway and are shown in order of significance (P values) listed on the right of the bars.

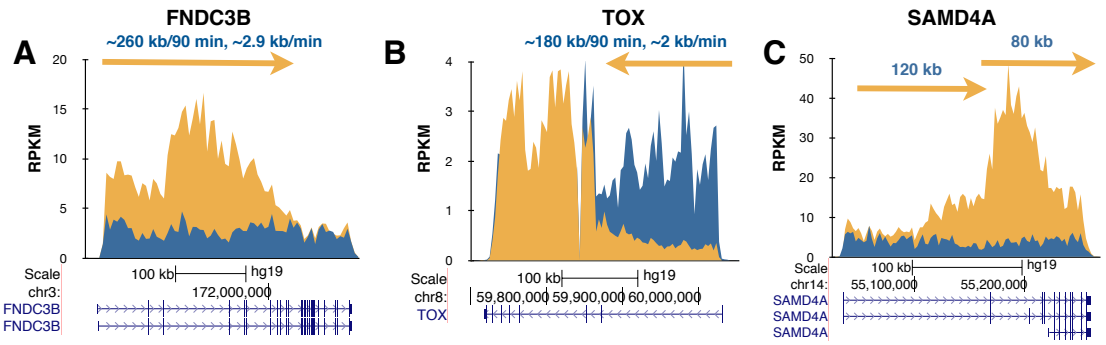


Figure S3.1: Bru-Seq analysis of TNF-treated human fibroblasts showing a wave of TNF-mediated induction and repression of tree large genes. (A) The FNDC3B gene is induced rapidly after exposure to TNF but since it is around 300 kb long, the "wave" of induced nascent RNA synthesis did not reach the end of the gene during the 90 minute treatment. (B) The synthesis of the TOX gene is rapidly inhibited by TNF exposure and the wave of reduced nascent RNA synthesis reached about 180 kb into the gene during the 90 minute incubation. (C) The SAMD4A gene is induced transiently by TNF treatment generating a wave of induced RNA synthesis with the rate of initiation returning to background levels by the end of the 90 min treatment. Blue color represent control and yellow represent a 60+30 min treatment with TNF. The gene maps are from RefSeq Genes (UCSC genome browser)

CHAPTER IV

Characterization of active promoters and enhancers in nascent RNA sequencing using BruUV-seq

4.1 Abstract

We present BruUV-seq, which utilizes UV light to introduce transcription-blocking DNA lesions prior to bromouridine-labeling and deep sequencing of nascent RNA. By inhibiting transcription elongation, but not initiation, pre-treatment with UV light causes a redistribution of the mapped reads, markedly enhancing the signal just downstream of transcription start sites (TSSs) and around active enhancers. Using BruUV-seq we show that while the majority of expressed genes utilized a single transcription start site (TSS), other genes had up to 5 active TSSs. Some gene clusters were transcribed from a single TSS suggesting their organization into operons in human cells. Furthermore, BruUV-seq identified treatment-inducible enhancers with increased eRNA production concomitant with up-regulation of nearby treatment-induced genes. Taken together, BruUV-seq is a powerful new approach for making detailed comparisons of promoter and enhancer utilization genome-wide within and between cell types.

4.2 Introduction

The main steps in the transcriptional cycle are initiation, elongation and termination. Initiation occurs after recruitment of RNA polymerase II to promoter elements by transcription factors and by the aid of active enhancer elements (Lenhard et al., 2012; Spitz and Furlong, 2012). Many genes have multiple promoter elements regulated by different transcription factors and epigenetic chromatin marks that allow for specific RNA isoform expression in different tissues or under different environmental conditions (Sandelin et al., 2007; Lenhard et al., 2012; Sanyal et al., 2012; Core et al., 2012). Which promoter(s) are utilized by a particular gene in a particular state is an important predictor of its context-specific RNA and protein product(s), and methods are required that fully reveal this variety of gene states.

Alteration in gene expression is an important strategy for cells to respond to a particular stimulus or stress. While it takes some time for these changes to be manifested in the total, steady-state RNA, these changes are almost immediately reflected in the pool of nascent RNA. RNA sequencing approaches using nascent RNA as a starting material, such as global run-on sequencing (GRO-seq) (Core et al., 2008), precision run-on sequencing (PRO-seq) (Kwak et al., 2013) and bromouridine labeling and sequencing (Bru-seq) (Paulsen et al., 2013b), have been developed to specifically study this nascent RNA pool.

A number of strategies exist for identifying transcription start sites (TSS) and enhancer elements on a genome-wide scale, but none has been optimized for nascent RNA. One approach is ChIP-seq analysis of either transcription initiation complexes (Venters and Pugh, 2013) or specific histone modifications such as H3K4me1, H3K4me3 and H3K27ac1. ChIP-seq analysis on its own does not conclusively pro-

vide information on promoter or enhancer usage but rather indicates the potential of a chromatin region or transcription initiation complex to mark a TSS or enhancer element. Another approach to identify TSS and enhancers genome-wide involves tagging of the 5'-ends of transcripts, examples of which include "cap analysis gene expression" (CAGE) (Shiraki et al., 2003; Valen et al., 2009), "5'-end serial analysis of gene expression" (SAGE) (Hashimoto et al., 2004) and "gene identification signature" analysis (GIS) with "pair end tags" (PET) (Ng et al., 2005). These techniques are very accurate in providing single-nucleotide resolution of TSSs, however, since these approaches rely on steady-state RNA they cannot account for alterations due to post-transcriptional processing. Furthermore, it has been reported that some recapping of degradation products of mature forms of RNA may occur in the cytoplasm and thus, some false TSS and enhancer predictions may occur when using CAGE-based assays (Lenhard et al., 2012).

RNA Pol II can generate RNA from enhancer elements, leading to the production of enhancer RNA (eRNA) whose function, if any, has not yet been established. eRNA is often transcribed from a bidirectional TSS and can be generated without polyadenylation (Kim et al., 2010). The chromatin surrounding enhancer elements is generally characterized by high levels of H3K4me1 and H3K27ac modifications and binding of acetyltransferase P30 while having low levels of H3K4me3 modifications (Zentner and Scacheri, 2012). Genome-wide annotations based on specific histone modifications have allowed for the identification of thousands of putative enhancer elements (Hoffman et al., 2013). However, not all putative enhancers generate eRNA. The ones that do often score higher in in vitro assays for enhancer activity suggesting that production of eRNA is linked to functional activity (Andersson et al., 2014).

Here we present BruUV-seq, an approach that complements the Bru-seq tech-

nique by enhancing nascent RNA signal around promoters and enhancers genome-wide. UVC light (254 nm) introduces predominantly cyclobutane pyrimidine dimers and 6-4 photoproducts in DNA that are distributed more or less randomly in the genome (Friedberg et al., 2006). These lesions are strong blocks to RNA polymerase II elongation complexes, causing them to stall (Donahue et al., 1994; Tornaletti and Hanawalt, 1999). In BruUV-seq, such lesions are introduced by UV irradiation prior to the metabolic labeling of nascent RNA with bromouridine (Bru), isolation of Bru-RNA and deep sequencing (Fig. 4.1a). Although elongating RNA polymerases stall at UV-induced lesions within gene bodies, new initiation and transcription near active TSS and enhancer elements is expected to continue (Donahue et al., 1994; Tornaletti and Hanawalt, 1999). The net result is an increase in read density at TSSs and at enhancer elements generating eRNA. BruUV-seq can thus readily determine how many promoters a particular gene is utilizing, identifying active enhancer elements genome-wide, and accurately determine changes in transcription levels both at promoter and enhancer regions following exposure to a stimulus or stress.

4.3 Results

4.3.1 UV light blocks elongation and redistributes RNA reads to TSSs

It is predicted that inhibition of gene expression by UV light is proportional to both the dose of UV light and the size of the gene. Inactivation of nascent RNA synthesis by randomly introduced transcription-blocking lesions by UV light has been previously used to determine genomic sizes of individual genes (Sauerbier and Hercules, 1978). To test what effect gene size has on the inactivation of nascent RNA synthesis on a genome-wide scale, we mock-irradiated or irradiated K562 cells with 20 J/m^2 of UVC light (254 nm) and immediately labeled nascent RNA with 2 mM bromouridine (Bru) for 30 min. Bru-labeled RNA was isolated and subjected to deep

sequencing (see Methods and Fig. 4.1a). A strong negative correlation was observed between the ratio of UV-irradiated to mock-irradiated RPKM values and gene size (Fig. 4.1b). Thus, the larger the gene, the less overall signal it gives following UV-irradiation. We also found that, similar to Bru-seq (Paulsen et al., 2013b), BruUV-seq was highly reproducible when performed in parallel on two similarly grown biological samples (Pearsons r=0.9971, Fig 4.1c).

We next examined the effect of UV-irradiation on the distribution of RNA reads within genes. Genes of at least 50 kb in length were aligned by their annotated TSSs and an aggregate view (median binned RPKM) was produced. Nascent Bru-seq data exhibited a relatively even distribution of signal from the TSSs into the gene (Fig. 4.1d). Following exposure to 25 J/m^2 (Fig. 4.1e) or 100 J/m^2 (Fig. 4.1f) of UVC radiation, the read distribution shifted markedly toward the TSSs in a dose-dependent manner. Although some of the enhancement of reads at some TSSs may be caused by UV-stimulated transcription initiation, the nearly uniform redistribution of reads across all genes in the genome indicates that the effect is strongly predominated by reduced generation of nascent RNA in gene bodies due to the UV-induced elongation blockage. The result is a relative rather than an absolute increase in read recovery near TSSs. Thus, we predicted that BruUV-seq could be used to reveal TSS utilization based on the redistribution of nascent RNA reads in the bodies of genes following UV-irradiation

4.3.2 Identification of active TSSs using BruUV-seq

To explore the behavior of individual genes, we performed Bru-seq to obtain genome-wide transcription rates, BruUV-seq to assess TSS usage, and our previously described BruChase-seq technique to obtain the splicing pattern of 6-hour old RNA (Paulsen et al., 2013b,a). In genes such as TLE4, which has one putative TSS

as predicted by its single H3K4me3 peak (Fig. 4.2a), BruUV-seq generated one single peak at the proximal boundary of the Bru-seq signal span, demonstrating how the phenomenon of UV-induced read redistribution enhances the signal around the TSS of an individual gene. Together, Bru-seq, BruUV-seq and BruChase-seq complement each other to provide a comprehensive picture of many unique aspects of the regulation of the TLE4 gene and other genes.

The primary transcripts of miRNAs are poorly annotated because these transcripts are rapidly processed into mature miRNA sequences of around 22 nucleotides while the rest of the primary transcript is degraded. In Figure 4.2b it can be seen that Bru-seq records a long nascent transcript emanating some 80 kb upstream of the mature MIR138-1 DNA sequence. BruUV-seq generated a single peak coinciding with a single H3K4me3 peak, suggesting that the TSS of the primary transcript is indeed located 80 kb upstream of the MIR138-1 sequence. BruChase-seq confirmed that the primary miRNA transcript was unstable since very little primary transcript signal was detectable after a 6-hour chase.

Other genes demonstrate efficient detection of more complex and multiple TSSs. For the divergent genes COL4A1 and COL4A2, the BruUV-seq technique readily confirmed the presence two very closely spaced divergent TSSs in human fibroblasts (Fig. 4.2c). For genes such as RERE for which multiple potential TSSs are suggested by multiple H3K4me3 peaks, BruUV-seq generated peaks over all four of these sites in human fibroblasts (Fig. 4.2d) but not in K562 cells where only three TSSs were used (Fig. 4.2e). Moreover, the relative intensity of the individual BruUV-seq peaks differed between the cell lines suggesting that their regulation of TSS usage differs fundamentally. This observation was consistent with the much higher Bru-seq signal in the proximal end of the RERE gene in K562, but only with the addition of BruUV-

seq data could the pattern of differential promoter utilization be fully appreciated.

To determine the extent of TSS complexity genome-wide in a single sample, we scored TSS usage at all individual genes using BruUV-seq. We found that the majority of expressed genes (95%) in human fibroblasts had only one active TSS, but we identified many genes utilizing 2 to 5 TSSs (Fig. 4.2f) (5% with $>$ 1 TSS). Thus, differential promoter usage is a substantial contributor to the generation of gene isoforms even for a given cell type in a given cell state.

Cap CAGE is a powerful technique to identify TSSs by capturing RNA molecules via their 5'-CAP. However, existing CAGE data has been generated from steady-state levels of RNA and is therefore not directly comparable to the nascent RNA analysis of TSS utilization performed here using BruUV-seq. As can be seen, the two techniques do not fully correspond to each other with some genes showing a TSS peak for BruUV-seq but not for CAGE. It is possible that even though the data is collected from the same cell line, different growth conditions may have lead to differences in gene expression. Alternatively, due to fast turnover rates or lack of 5'-capping, some transcripts are not easily captured by CAGE in steady-state RNA isolations. For these reasons we did not further compare these techniques.

4.3.3 Potential operons in human cells

Gene clusters coding for related proteins transcribed in the same orientation sometimes share a common promoter and are transcribed as a "neighborhood" or operon. This polycistronic arrangement is common in bacteria and in some eukaryotes such as *C. elegans* and *Drosophila melangoster* (Spieth et al., 1993; Blumenthal, 2004). However, the evidence for the presence of operons in human cells is scarce. To identify potential operons in human cells, we compared mapped data from both Bru-seq and BruUV-seq. We first analyzed a gene cluster coding for zinc finger proteins on

chromosome 12 and although these genes were transcribed at similar levels and in the same orientation, BruUV-seq revealed that each of the genes utilized their own TSS (Fig. 4.3a). In contrast, a region on chromosome 14 encoding a large set of snoRNAs and miRNAs appeared to be transcribed as a neighborhood from a single common TSS (Fig. 4.3b). Furthermore, the TTTY15 and USP9Y genes on the Y chromosome appeared to be expressed from a common TSS according to the BruUV-seq data and the BruChase-seq data shows that the primary transcript appeared to have been spliced within the 6-hour chase (Fig. 4.3c). Taken together, the BruUV-seq approach clarifies where genes initiate transcription, whether multiple promoters are used and, in combination with Bru-seq, whether genes may be organized into neighborhoods utilizing a common promoter.

4.3.4 Use of BruUV-seq to validate gene fusions

Gene fusions are the products of aberrant recombination of normally separate genes. Many identified gene fusions are oncogenes known to cause or contribute to cancer (Rowley, 1973; Kumar-Sinha et al., 2006). Chronic myelogenous leukemia (CML) is caused by a chromosomal translocation between chromosomes 9 and 22 that fuses the 5' portion of the BCR gene to the 3' portion of the ABL1 gene29. In K562 cells, a CML-derived cell line known to harbor BCR-ABL, Bru-seq revealed high BCR expression through most of the 5'-part of the gene with a drastic drop in RNA reads at the known translocation site at the 3'-end of the gene. The ABL1 gene, on the other hand, exhibited low expression at its 5'-end, which increased dramatically about 10 kb into the gene. In the absence of prior knowledge of a gene fusion at this locus, the abrupt increase in signal from the ABL1 gene could be interpreted as initiation from a strong, un-annotated promoter located in the first intron. However, BruUV-seq showed no UV-induced peak to support this notion,

but did identify the strong active BCR TSS known to drive BCR-ABL expression. Thus, BruUV-seq can be used to clarify and validate the existence of gene fusions in cells, as well as to predict the location of patient-specific translocation junctions.

4.3.5 BruUV-seq and Bru-seq signals are positively correlated

As discussed above, gene size greatly affects the relationship between gene expression when measured by Bru-seq or BruUV-seq (Figure 4.1b). We reasoned that by using only short regions downstream of the TSS for measuring expression, this gene size bias would be eliminated. Indeed, the ratio between Bru-seq and BruUV-seq signal in the first 5 kb of genes was not correlated to the sizes of the genes. Furthermore, the Bru-seq and BruUV-seq expression measurements within the first 5 kb of the genes were strongly correlated. The BruUV-seq signal in the first 5 kb after the TSS was also correlated with Bru-seq expression along the whole length of the genes, albeit more weakly. This difference was more pronounced for genes expressing two or more TSS. The positive correlation between the two techniques suggests that BruUV-seq data are predictive of Bru-seq data and therefore BruUV-seq could be used as a surrogate for nascent RNA transcription measurements with Bru-seq.

Using a restricted portion of a gene to calculate its expression value, however, is counterintuitive since one generally wants to use all available data. To determine if restricting the expression measurement was reasonable, we compared the amount of variation observed in both approaches. We randomly selected a given number of reads 10 times from both Bru-seq and BruUV-seq samples to simulate a given read depth. The median coefficient of variation in gene expression from Bru-seq and BruUV-seq were very similar, but consistently lower in Bru-seq, regardless of the simulated read depth assessed. Interestingly, higher doses of UV (100 J/m^2) lead to smaller overall coefficients of variation when compared to lower doses of UV (25

J/m^2). This is likely caused by the greater accumulation of mapped reads within the first 5 kb downstream of the TSS at higher UV doses (Fig. 4.1e&f) and suggests that BruUV-seq performs better at higher doses of UV.

4.3.6 Using BruUV-seq to assess induced initiation of transcription

We recently used Bru-seq to investigate transcriptional changes caused by TNF-mediated induction of the acute inflammatory response (Paulsen et al., 2013b). Here we explored whether we could use BruUV-seq to assess changes in gene expression following TNF treatment by measuring the signal in the first 5 kb of genes. NFKB1 is a gene that is induced 9.7-fold by TNF as assessed by Bru-seq (Fig. 4.4a). Using BruUV-seq to assess the fold difference in transcription reads over the first 5 kb between control and TNF treated cells we observed a 7.8-fold difference. The LBH gene was found to be down-regulated 5.2-fold by TNF as measured by Bru-seq and 4.7-fold when measured over the first 5 kb using BruUV-seq (Fig. 4.4b). The ARHGAP24 gene showed an isoform-specific TNF-induction (Fig. 4.4c) illustrating the usefulness of BruUV-seq to determine isoform-specific regulation of initiation. Genome-wide comparisons of TNF-mediated changes in transcription between Bru-seq and BruUV-seq yielded a strong correlation (Fig 4.4d). Thus, measuring changes in transcription of the first 5 kb of genes using BruUV-seq can be used as a reasonable surrogate for measuring changes across the whole gene with Bru-seq. This is important for large genes, which take more time to complete synthesis following an acute treatment. In such cases, BruUV-seq captures the induced changes in transcription initiation (and early elongation) more accurately, changes that might be missed by techniques restricted to whole genes or mature mRNAs.

4.3.7 UV light increases read density at putative enhancer elements

It has been shown that RNA can be generated from certain putative enhancer elements and this RNA has been termed eRNA (Kim et al., 2010). It is not well understood how the eRNA is produced and what function it may have in regulating genes. The fact that the eRNA is capped and the similarities between enhancers and promoters depleted of CpG islands (Andersson et al., 2014) would suggest that active enhancers may function as transcription initiation sites. Assuming that transcription initiates from certain enhancers, one would expect to observe increased BruUV-seq signal around them. Indeed, in addition to increased read density at TSSs of genes, BruUV-seq showed enhancement of RNA reads in narrow peaks in intergenic regions. One of these peaks was located upstream of the FOS gene in a region of a well-known enhancer element (Fig. 4.5a). The fold enhancement of reads in this region obtained by BruUV-seq compared with Bru-seq appeared to be much greater than could be accounted for by simple redistribution of reads from areas of inhibited elongation to areas downstream of TSSs. It is possible that the extremely unstable eRNA is stabilized when associated with RNA polymerases stalled at UV-induced lesions. In support of this are findings that Bru-labeled RNA was turned over much more slowly if the cells were UV-irradiated after the Bru-labeling. Interestingly, certain genomic regions had a high concentration of BruUV-seq peaks that aligned with the enhancer signatures of high levels of H3K4me1 and H3K27ac and a low level of H3K4me3. The area in Figure 4.5b is one of such regions that appears as an "enhancer forest" and is located at a similar genomic position (upstream from THBS1) to a so called super-enhancer found in mice macrophages (Whyte et al., 2013).

In order to determine if this increased signal around enhancers happened in a genome-wide fashion, the ENCODE project's combined genome segmentation anno-

tation was used (Hoffman et al., 2013). An aggregate view of the reads surrounding the intergenic enhancer regions demonstrated that the signal was bidirectional, as expected for eRNA (Fig. 4.5c) (Kim et al., 2010). Accumulation of sequencing reads in intergenic genomic regions classified as enhancers was greater in BruUV-seq than in Bru-seq. Furthermore, the enhancement in eRNA signal was greater for the sample irradiated with the higher UV dose (100 J/m^2) than with the lower UV dose (25 J/m^2). Importantly, BruUV-seq demonstrated a greater amount of eRNA signal when compared to the ENCODE project's subcellular fractionation mature RNA libraries (ENCODE Project Consortium et al., 2012), suggesting that BruUV-seq improves the sensitivity of eRNA detection over other approaches (Fig. 4.5d).

4.3.8 Changes in gene expression are accompanied by changes in eRNA production

A positive correlation has been observed between the levels of eRNA production and the expression level of their closest gene (Kim et al., 2010). While it would be reasonable to expect changes in gene expression to be correlated with similar changes in eRNA production, this has not been conclusively shown. After TNF exposure of human fibroblasts HF1, we observed a sharp induction of NFKB1 gene expression as seen with Bru-seq (Fig. 4.5e) (Paulsen et al., 2013b). Using BruUV-seq we observed two strong peaks about 40-50 kb upstream of the TSS for the NFKB1 gene that aligned with the enhancer marks H3K4me1 and H3K27ac and the intensity of these two peaks increased dramatically following TNF treatment. A genome-wide approach for the identification of enhancers based on the ratio between BruUV-seq and Bru-seq signal was carried out. Briefly, intergenic regions with BruUV-seq signal enhancement (UVE regions) were identified using a Hidden Markov Model. The TNF-induced change in eRNA expression in these UVE regions was correlated to the TNF-induced change in pre-mRNA expression of their nearest gene. A positive

correlation was observed (R=0.428), which suggests that changes in gene expression and eRNA production are correlated genome-wide (Fig. 4.5f). Taken together, BruUV-seq is a powerful new technique to identify both constitutive and inducible enhancer elements genome-wide.

4.4 Discussion

Here we present BruUV-seq as a companion technique to Bru-seq where cells are UV-irradiated prior to metabolic labeling of nascent RNA (Fig. 4.1a). Due to inhibition of transcription elongation, but not transcription initiation, UV pretreatment redistributes the bromouridine labeling of nascent RNA toward the beginning of transcription units, and thus TSSs. In addition, ordinarily unstable RNAs appear to be protected from degradation when persistently bound to stalled RNA polymerases, leading to the markedly increased yields of reads corresponding to eRNAs. These effects of UV light form the basis of the BruUV-seq approach and allow for the identification of active TSSs and enhancer elements genome-wide.

The analysis of Bru-seq and BruUV-seq data in parallel gives novel insight into nascent transcription and therefore context-dependent gene function. We demonstrate on a genomic scale that transcription inhibition by UV is related to gene size, confirming the findings by Sauerbier and Hercules (1978) when estimating genomic sizes of individual genes (Fig. 4.1b). It has been proposed that UV light may cause inhibition of transcription initiation (Rockx et al., 2000) but in contrast, our data suggest that transcription elongation is the primary target of UV-mediated inactivation of transcription. We were able to identify thousands of active TSSs genome-wide and hundreds of the genes examined presented more than one active TSS giving rise to distinct isoforms (Fig. 4.2). We found examples of gene clusters driven by indi-

vidual promoters (Fig. 4.3a) as well as gene clusters driven by a single TSS (Figs. 4.3b&c). Furthermore, we detected multi-TSS containing genes where the different TSSs responded differently to a specific stimulus (e.g. Fig. 4.4c). In the K562 cell line, BruUV-Seq confirmed the presence of the BCR-ABL gene fusion by showing that this fusion gene initiated from a single TSS in the BCR gene.

In addition to identifying TSS usage, we found that BruUV-seq data from the first 5 kb of genes could be used to infer transcription levels of the entire gene. This application of BruUV-seq proved especially important for very large genes or when genes harbor multiple active TSS, since each TSS can be identified and analyzed independently without having to wait for mature RNAs to be formed. BruUV-seq also enabled us to observe treatment-induced changes in eRNA production (Fig. 4.5e). The positive correlation observed in treatment-induced changes in eRNA and pre-mRNA is extremely important since it might indicate that these in silico identified sites are in fact regulators of the selected genes. Much effort has been given to better understand the relationship between regulatory regions and gene transcription using tools such as DNAse Hypersensitivity Sites (Thurman et al., 2012), ChIA-PET (Li et al., 2012) and CAGE (Andersson et al., 2014). We believe that the abilities of BruUV-seq to assess nascent transcription and to enrich for nascent RNA signal in promoter and enhancer regions makes it a very powerful technique to explore the mechanisms of gene regulation genome-wide.

4.5 Online Methods

4.5.1 Cell culturing

Human diploid foreskin fibroblasts HF1 (a gift from Mary Davis, University of Michigan) expressing hTERT (Ljungman and Zhang, 1996; Paulsen et al., 2013b,a; Veloso et al., 2013, 2014) were grown as monolayers in MEM supplied with 10% fetal

bovine serum and antibiotics (Invitrogen). K562 cells were grown in suspension in IMDM with 10% FBS.

4.5.2 UV-irradiation and bromouridine labeling of cells

The media of adherent cells grown on 100 mm plates were removed and 100 μl of PBS was added to keep the cells from drying out during the UV-irradiation. Suspension cells were gently pelleted and suspended in 1 ml of PBS and placed in a 100 mm plates for UV-irradiation at room temperature. Cells were irradiated in 100 mm plates without the lid on with different doses of 254 nm UVC light. The irradiation source (Philips, New York, NY) generated UVC light with a dose rate of 1 $J/m^2/s$ as measured with a UVX radiometer (UVP, Inc. Upland CA). Immediately following UVC irradiation, the cells were supplied with conditioned media containing 2 mM bromouridine (BrU) (Aldrich) and they were incubated for 30 minutes to label nascent RNA. Isolation of Bru-containing RNA, cDNA library preparations and deep sequencing were performed as previously described (Paulsen et al., 2013b,a).

4.5.3 Read mapping and gene annotations

The sequenced reads were initially mapped to human ribosomal DNA complete repeating unit (U13369.1). Reads that remained unaligned were mapped to the human genome hg19 build (Paulsen et al., 2013b,a). The RefSeq annotated isoforms were merged to create a simplified annotation with one entry for each gene. This simplified annotation was used for most analysis in this manuscript. The identification of multiple TSS usage in genes was carried out using the Ensembl gene annotation (release 69) (Flicek et al., 2013). Isoforms for a gene with a TSS within 1 kb of each other were merged into a single isoform, and the most upstream TSS was used.

4.5.4 Identification of active TSSs

In order to identify active TSSs, a two state Hidden Markov Model (HMM) similar to the one described by Veloso et al. (2014) was used. The goal was to determine if a peak in the BruUV-seq signal occurred within 500 bp from the annotated TSS. For this, the HMM model attempted to recognize a state prior to the TSS (state 1), characterized by low expression values, and a state after the TSS (state 2), characterized by a large increase in signal. The genomic region from 5 kb upstream to 5 kb downstream of the TSS was split into 250 bp bins. The expression signal within each bin was measured for the BruUV-seq samples. The expression signal was initially quantile normalized. Next, a z-score Gamma-equivalent normalization of the data was carried out. This normalized data was used as the observed output of the bins. The signal from two sections, 5 kb-2 kb upstream from the TSS and 2 kb-5 kb downstream from the TSS, were used to determine the emission probabilities of state 1 and 2 consecutively. The only possible transition between states was from 1 to 2, and its probability was set to 0.00001. The model was fit to the data and the Viterbi algorithm was used to determine the most likely state of each bin using the R package msm (Jackson, 2011; Maechler et al., 2013). A TSS was considered to be active if a transition from state 1 to state 2 occurred within 500 bp of the annotated TSS.

4.5.5 ENCODE RNA-seq data

The ENCODE's long RNA-seq raw reads (ENCODE Project Consortium et al., 2012) were downloaded from `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCshlLongRnaSeq`. In order to carry out a direct comparison between Bru-seq, BruUV-seq and the ENCODE's RNA-seq data, the ENCODE's

RNA-seq reads were trimmed to 52 base pairs and remapped as single-reads using the pipeline described in (Paulsen et al., 2013b,a). The following K562 samples were used: whole cell PolyA-minus replicates 1 and 2 (GEO ID: GSM758577); whole cell PolyA-plus replicates 1 and 2 (GEO ID: GSM765405); chromatin bound replicates 3 and 4 (GEO ID: GSM765392); nucleoplasm replicates 3 and 4 (GEO ID: GSM765390); nucleus PolyA-minus replicates 1 and 2 (GEO ID: GSM767844); nucleus PolyA-plus replicates 1 and 2 (GEO ID: GSM765387).

4.5.6 Determining eRNA expression

Two different approaches were taken to determine genomic regions that could represent enhancers. The first approach used the ENCODE's combined genome segmentation data (Hoffman et al., 2013). Only segments determined to be enhancers (class E) were used in the analysis. In the second approach, we carried a de novo discovery of regions with BruUV-seq signal enhancement (see 4.5.7). Since we were interested in measuring eRNA expression, it was very important to avoid mRNA producing sites. Therefore, only intergenic sites were used. Intergenic sites were defined as regions that did not overlap genes or their transcription units (defined in (Paulsen et al., 2013b)). The signal within these sites was used to determine the expression rate, measured in RPKM.

4.5.7 Identification of UV enhancement peaks

Relative to Bru-seq, BruUV-seq results in an increase in RPKM values just downstream of a TSS and a decrease at more distal gene positions (e.g. Fig. 4.2a), which led us to develop a genome-wide Hidden Markov model with two states: UV enhanced (UVE) and UV repressed (UVR). Read counts for paired Bru-seq and BruUV-seq samples were first independently aggregated into 250 bp bins and subjected to wavelet

smoothing. Paired data were then normalized to a common scale based on total sample read counts and the fraction of UV reads (f_{UV}) determined for each bin, where each read was considered to be a Bernoulli trial with a sample-pair-dependent probability of being from the BruUV-seq sample. Bins in annotated genes with a Bru-Seq gene RPKM of at least 0.25 that were more than 20 Kb downstream of the TSS were then used to determine the cumulative f_{UV} in the presumptive UVR portion of each gene. The mean and variance of these f_{UV} values were used to calculate the α and β shape parameters of the binomial distribution with overdispersion of f_{UV}, i.e. the beta binomial distribution, for the UVR state. Similar shape parameters were obtained for the UVE state by reflecting the UVR beta distribution such that bins with a high probability of being UVR had a low probability of being UVE. Emissions probabilities for the UVR and UVE states were then calculated for each bin using these shape parameters and the bin's normalized Bru-seq and BruUV-seq read counts. The Viterbi algorithm was finally solved across all genome bins using a fixed transition probability of 0.005, with sequences of UVE bins taken as UVE peaks. Importantly, this process only used gene annotations to train the model, and so could detect a UVE peak at any genome location.

4.6 Acknowledgements

We thank the personnel at University of Michigan Sequencing Core for professional technical assistance and Manhong Dai and Fan Meng for administration and maintenance of the University of Michigan Molecular and Behavioral Neuroscience Institute (MBNI) computing cluster. This work was supported by pilot funding from University of Michigan Bioinformatics Program, the University of Michigan Biomedical Research Council, the Will and Jeanne Caldwell Endowed Research Fund of the

University of Michigan Comprehensive Cancer Center, University of Michigan School of Public Health (NIEHS P30 pilot grant), Department of Defense, Uniting Against Lung Cancer, National Institute of Environmental Health Sciences (1R21ES020946), National Human Genome Research Institute (1R01HG006786) and the National Institute of Health (P50CA130810).

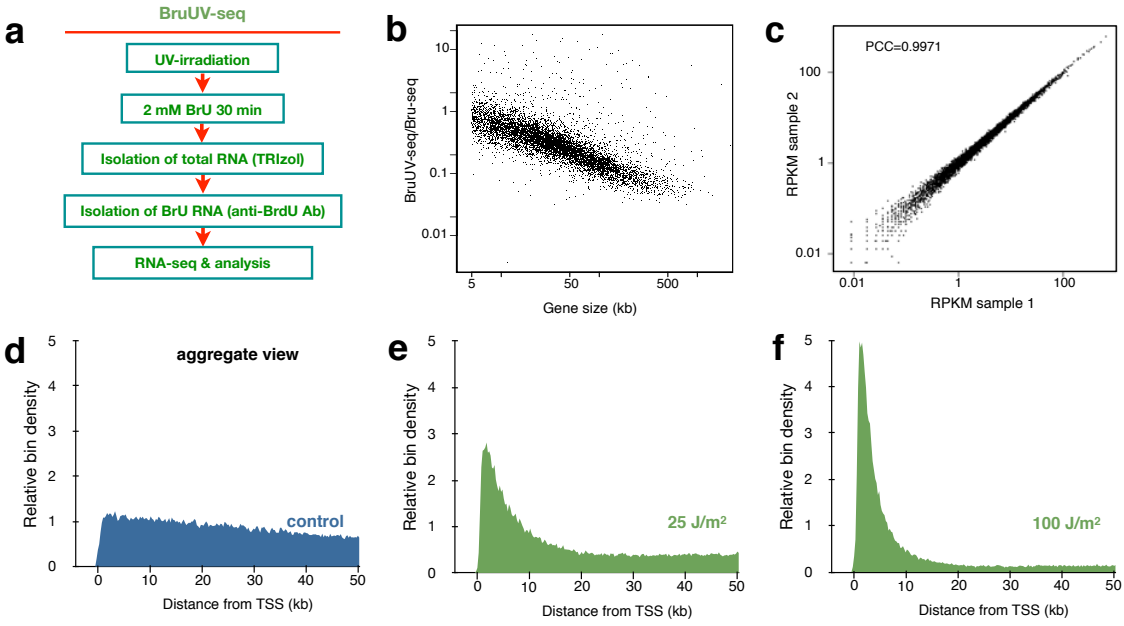


Figure 4.1: Comparison of Bru-seq and BruUv-seq signal. (a) Experimental outline of the BruUV-Seq technique. (b) Strong correlation between UV-mediated inhibition of transcription and gene size. The ratio of integrated transcription reads over whole genes following UV-irradiation in HF1 human fibroblasts compared to mock-irradiated HF1 cells are shown on the Y-axis while gene size is shown on the X-axis. (c) Reproducibility of BruUV-seq. The RPKM values of the first 5 kb of genes expressed above 0.5 RPKM were compared between two biological experiments involving human HF1 fibroblasts in which cells had been UV-irradiated, Bru-labelled and sequenced in parallel. The Pearson's correlation coefficient (PCC) between the two replicas was 0.9971. (d) BruUV-seq data from K562 cells that had been mock-irradiated. (e), irradiated with $25J/m^2$ or (f) with $100J/m^2$. The transcription reads from genes with a length of 50 kb or larger were compiled in an aggregate view where all the genes had been lined up from their TSS and expressed as relative bin density.

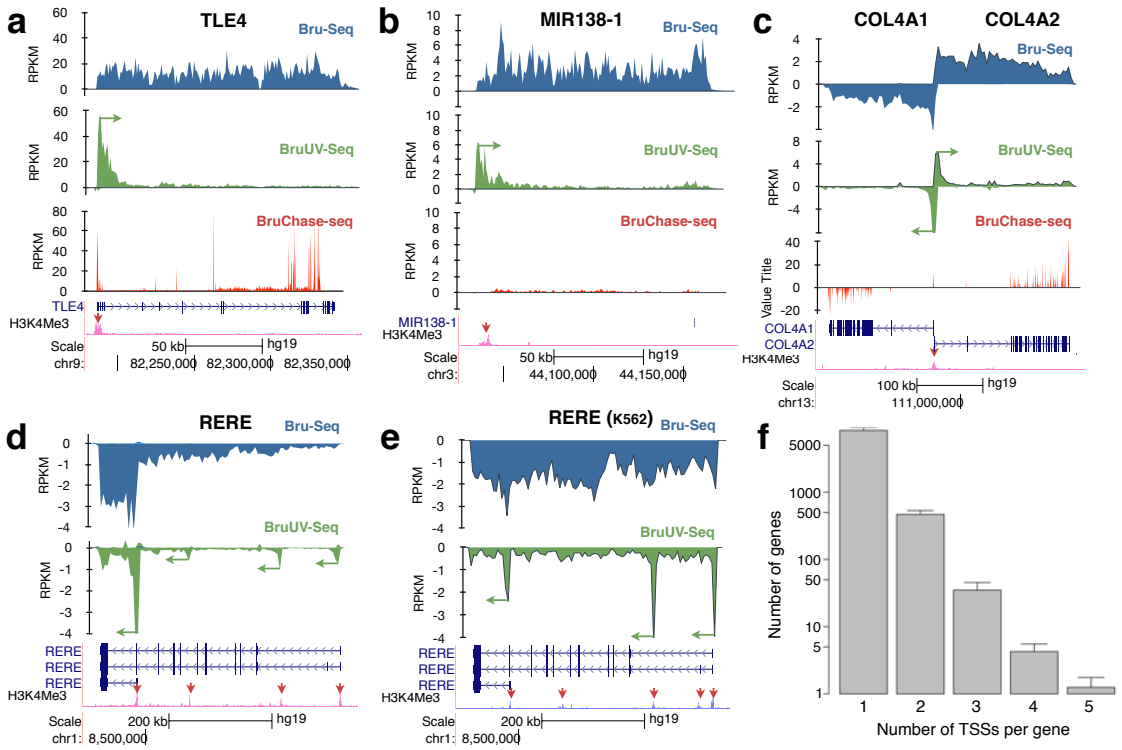


Figure 4.2: BruUV-seq identifies TSSs genome-wide. (a) BruUV-seq verifies that the TLEA4 gene in HF1 cells is transcribed from a single TSS that aligns with H3K4me3 marks (red arrow). (b) The primary transcript of MIR138-1 is initiating transcription from a single TSS about 80 kb upstream of the MIR138-1 sequence. (c) The COL4A1 and COL4A2 genes are transcribed from two divergent TSSs. (d) BruUV-seq identifies four distinct TSSs in the RERE gene in human fibroblasts but only three in K562 cells (e). (f) Genome-wide assessment of the number of TSSs per gene with BruUV-seq. Data is taken from three independent biological experiments using HF1 cells and shows that many genes have multiple TSSs.

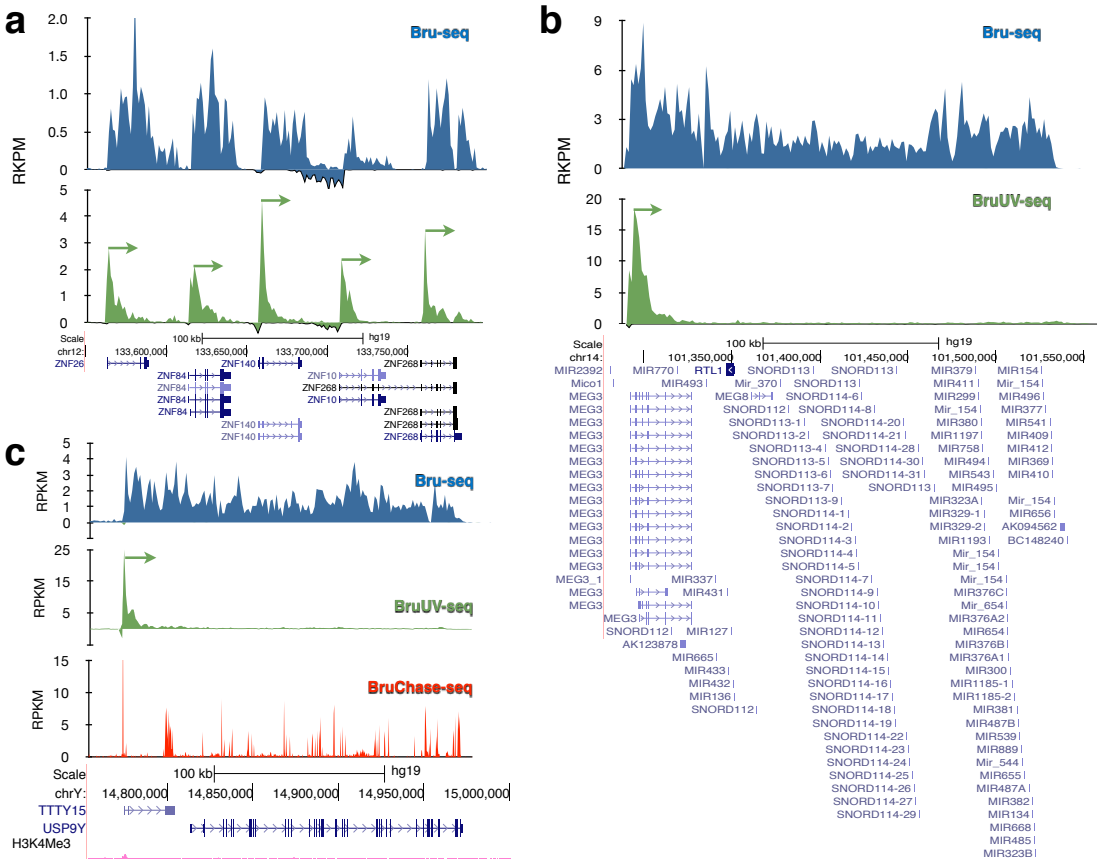


Figure 4.3: BruUV-seq distinguishes between gene clusters initiating from individual or common promoters in HF1 cells. (a) Bru-seq (blue) and BruUV-seq (green) data of genes encoding the zinc-finger proteins ZNF26, ZNF84, ZNF140, ZNF10 and ZNF268 on chromosome 12 are shown. Each gene uses their individual TSS to direct their transcription. (b) Bru-seq and BruUV-seq data of a gene cluster on chromosome 14 transcribing as an operon from a common TSS. (c) Bru-seq, BruUV-seq and BruChase (6 h chase) data for the TTTY15 and USP9Y genes transcribing as an operon from a common TSS. Note that the primary transcript appears to be spliced. The gene maps are from RefSeq taken from the UCSC web browser.

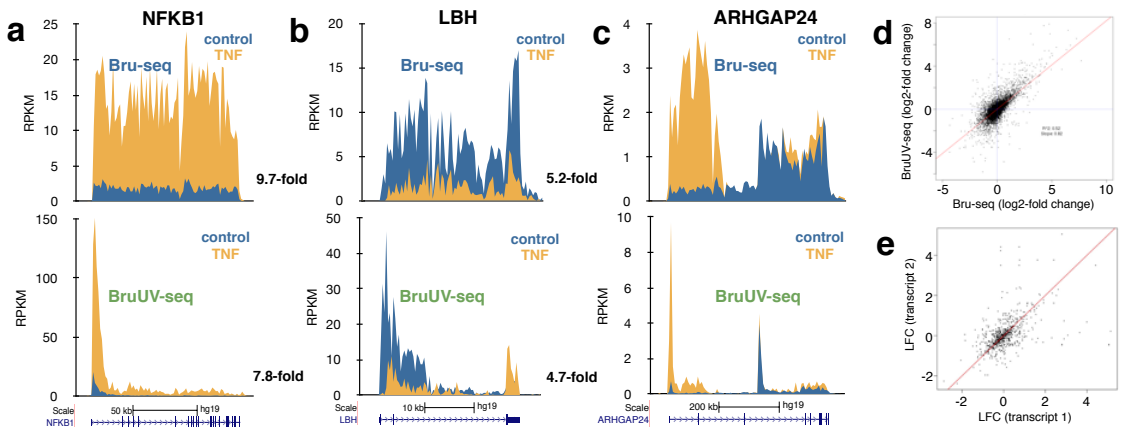


Figure 4.4: BruUV-seq can be used to predict Bru-seq data for gene induction. (a) Top: The NFKB1 gene is induced 9.7-fold following treatment with TNF in human fibroblasts HF1 as measured with Bru-seq (see Paulsen et al. (2013b) for details). Bottom: Comparing the read density over the first 5 kb of the NFKB1 gene of control and TNF-treated cells, we found that BruUV-seq data yield a 7.8-fold enrichement by TNF which is similar to the 9.7-fold induction measured with Bru-seq at top. (b) Top: Transcription from the LBH gene is reduced 5.2-fold measured with Bru-seq and 4.7-fold measured over the first 5 kb using BruUV-seq at bottom. (c) Isoform-specific induction of the ARHGAP24 gene shown with Bru-seq data on top is confirmed with BruUV-seq at bottom. (d). Genome-wide analysis of the correlation between TNF-induced changes in gene expression using the full gene in Bru-seq (x-axis) and the first 5 kb of the gene in BruUV-seq (y-axis). (E) Pairwise comparison of the TNF-induced log-fold change in gene expression of transcripts in genes with multiple active TSS. Points distant from the identity line (red line) indicate genes with TSS-specific response to TNF.

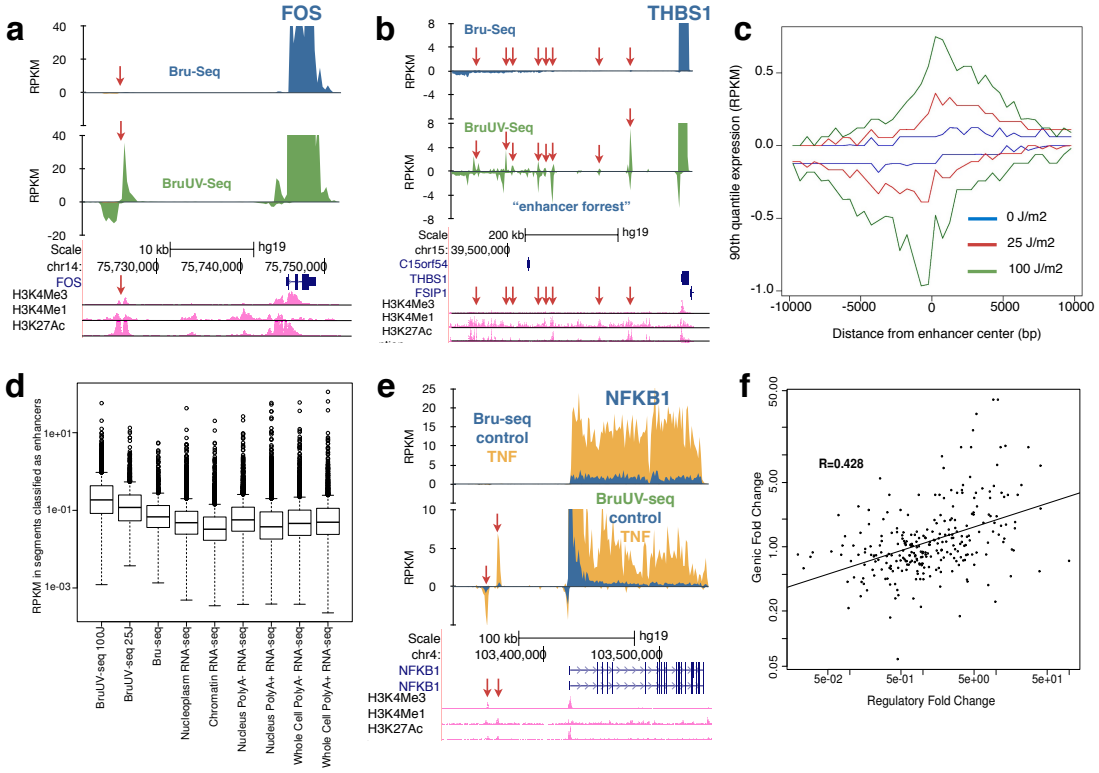


Figure 4.5: Use of BruUV-seq to identify active enhancer elements genome-wide. (a) Bru-seq (blue) and BruUV-seq (green) data for the FOS gene and upstream enhancer element (red arrow) in HF1 cells. (b) Bru-seq and BruUV-seq data for a "enhancer forest" region (red arrows) upstream of the highly expressed THBS1 gene in HF1 cells. (c) An aggregate view of the reads surrounding 526 intergenic enhancer regions as defined by the ENCODE project's combined genome segmentation annotation. It can be seen that the aggregate signal of eRNA was bidirectional and the enhancement of reads was proportional to the UV dose. (d) Boxplot indicating distribution of expression values (RPKM) observed in intergenic ENCODE defined enhancer regions in K562. Data includes BruUV-seq, Bru-seq and ENCODE's cellular partitioning long RNA-seq samples. (e) Comparison of Bru-seq (top) and BruUV-seq (bottom) for the NFKB1 gene and upstream region in HF1 cells before and after treatment with TNF for 1 hour. Enhancer elements activated by TNF are shown with red arrows. (f) Correlation between TNF-induced changes in eRNA and mRNA expression in human fibroblasts. Enhancers were identified using a Hidden Markov Model based on the BruUV-seq and Bru-seq expression profiles and were assigned to their nearest genes.

CHAPTER V

Genome-Wide Transcriptional Effects of the Anti-Cancer Agent Camptothecin

5.1 Abstract

The anti-cancer drug camptothecin inhibits replication and transcription by trapping DNA topoisomerase I (Top1) covalently to DNA in a "cleavable complex". To examine the effects of camptothecin on RNA synthesis genome-wide we used Bru-Seq and show that camptothecin treatment primarily affected transcription elongation. We also observed that camptothecin increased RNA reads past transcription termination sites as well as at enhancer elements. Following removal of camptothecin, transcription spread as a wave from the 5'-end of genes with no recovery of transcription apparent from RNA polymerases stalled in the body of genes. As a result, camptothecin preferentially inhibited the expression of large genes such as proto-oncogenes, and anti-apoptotic genes while smaller ribosomal protein genes, pro-apoptotic genes and p53 target genes showed relative higher expression. Cockayne syndrome group B fibroblasts (CS-B), which are defective in transcription-coupled repair (TCR), showed an RNA synthesis recovery profile similar to normal fibroblasts suggesting that TCR is not involved in the repair of or RNA synthesis recovery from

Official citation:
Veloso, A., Biewen, B., Paulsen, M.T., Berg, N., Carmo de Andrade Lima, L., Prasad, J., Bedi, K., Magnuson, B., Wilson, T.E., Ljungman, M. Genome-Wide Transcriptional Effects of the Anti-Cancer Agent Camptothecin. *PLoS ONE*, 2013 8(10): e78190. doi:10.1371/journal.pone.0078190

transcription-blocking Top1 lesions. These findings of the effects of camptothecin on transcription have important implications for its anti-cancer activities and may aid in the design of improved combinatorial treatments involving Top1 poisons.

5.2 Introduction

DNA topoisomerase I (Top1) relaxes torsional tension that is generated in the DNA helix as a consequence of replication, transcription and chromatin remodeling (Liu and Wang, 1987; Pommier, 2006). The Top1-mediated reaction involves the covalent binding to DNA, cleavage of one strand of the DNA helix followed by the passing of the other strand through the break and finally the resealing of the DNA strand break. The anti-cancer drug camptothecin specifically inhibits Top1 (Hsiang and Liu, 1988) by acting prior to the resealing step, effectively trapping Top1 covalently bound to the DNA in a "cleavable complex". Camptothecin and other Top1 poisons are used for the treatment of ovarian, cervical, colon, pancreatic, lung, breast, prostate and brain cancers (Pommier, 2013). The anti-cancer activity of camptothecin is thought to be linked to replication-mediated toxicity (Hsiang et al., 1989). However, the inhibitory effect of camptothecin on transcription has also been acknowledged to contribute to toxicity especially in non-dividing cells (Wu and Liu, 1997).

It has been shown that camptothecin-stabilized Top1-DNA complexes retard elongation but not initiation of transcription (Ljungman and Hanawalt, 1996). In response to transcription blockage, Top1 is targeted for degradation in an ubiquitin-dependent manner (Desai et al., 1997) and subsequent residual DNA-bound amino acid residues may require the action of tyrosyl-DNA phosphodiesterase 1 (TDP1) for their removal in order for transcription elongation to resume (Plo et al., 2003). Block-

age of the transcription machinery by Top1 complexes trapped on DNA by camptothecin have been shown to lead to the induction of DNA double strand breaks (Wu and Liu, 1997) and the formation of DNA-RNA hybrid structures (R-loops) activating the stress kinase ATM (Sordet et al., 2009; Sakasai et al., 2010). Furthermore, this transcription stress results in activation of the p53 response pathway (Ljungman et al., 1999, 2001; Ljungman and Lane, 2004; Lin et al., 2013) and induction of 53BP1-mediated DNA damage processing (Sakai et al., 2012). Top2 and PARP1 play overlapping roles to Top1 in non-dividing cells. Combination of Top1 and Top2 or PARP-targeting drugs may be effective in non-dividing tumor cells (Lin et al., 2013).

Following camptothecin reversal the topoisomerase reaction is completed and transcription complexes are thought to resume elongation. Interestingly, RNA synthesis recovery from the Dhfr gene in CHO cells following camptothecin removal was found to resume as a wave in a 5'-3' direction with no apparent recovery downstream in the gene suggesting that transcription complexes blocked by trapped Top1 complexes are unable to resume elongation in this gene following camptothecin reversal (Ljungman and Hanawalt, 1996). Cells derived from patients with Cockayne's syndrome (CS) are hypersensitive to CPT due to more double strand breaks induced during S-phase (Squires et al., 1993). The CSB protein has been suggested to be involved in the repair of covalently DNA-linked Top1 (Horibata et al., 2011) and this may explain why the recovery of total RNA synthesis slower in CS cells (Squires et al., 1993; Horibata et al., 2011; Lin et al., 2013). However, other studies have found no defect in RNA synthesis recovery following CSB knockdown (Sakai et al., 2012).

In this study, we explored the effects of camptothecin on RNA synthesis genome-wide using Bru-Seq. This technique is based on the metabolic labeling of RNA

using bromouridine (Bru) followed by specific isolation of Bru-labeled nascent RNA, library preparation and deep sequencing Paulsen et al. (2013b). Our results show that the Top1 inhibitor camptothecin affects many aspects of transcription where blockage of transcription elongation combined with the apparent lack of recovery of synthesis from RNA polymerases blocked in the body of the genes cases a preferential inhibition of expression of large genes. Furthermore, we find no defect in RNA synthesis recovery in CS-B cells following camptothecin reversal suggesting that TCR may not be required for this recovery.

5.3 Materials and Methods

5.3.1 Cell lines, camptothecin treatment and Bru-Seq

hTERT immortalized diploid human foreskin fibroblasts (gift from Dr. Mary Davis, Department of Radiation Oncology, University of Michigan) (Ljungman and Zhang, 1996; Ljungman et al., 1999) and CS-B fibroblasts (GM00739, Coriell Cell Repository) were grown as monolayers in MEM supplied with 10% fetal bovine serum and antibiotics (Invitrogen). Cells were treated for 45 min with $20\mu M$ camptothecin (Sigma) and labeled for 15 min with $2mM$ bromouridine either during the last 15 min of camptothecin treatment or following washout. The Bru-Seq and BruChase-Seq procedures were performed as previously described (Paulsen et al., 2013b). In short, total RNA was isolated from the cell samples using TRIzol reagent (Invitrogen) followed by specific isolation of Bru-labeled RNA using anti-BrdU antibodies (BD Biosciences) conjugated to magnetic beads (Dynabeads, Goat anti-Mouse IgG, Invitrogen). The isolated RNA was then converted into a strand-specific DNA library using the Illumina TruSeq Kit (Illumina) as previously described (Paulsen et al., 2013b).

5.3.2 Illumina Hi-Seq sequencing and data analysis

Sequencing of the cDNA libraries was performed by the staff at the University of Michigan Sequencing Core using the Illumina HiSeq 2000 sequencer. Base calling was performed using Illumina Casava v1.8.2. and read mapping was performed using TopHat, accepting only reads that could be mapped uniquely to the genome. We calculated RPKM values from the Bru-Seq data and plotted the data using a custom-built browser as previously described (Paulsen et al., 2013b).

5.3.3 Data availability

The primary data used in the analyses will be deposited at NCBI's Gene Expression Omnibus at the time of publication. We will upload the original genome mapping BAM files and the derived synthesis lists as BED files.

5.4 Results

5.4.1 Camptothecin preferentially inhibited RNA synthesis of large genes

We challenged human fibroblasts for 45 min with $20\mu M$ camptothecin and labeled RNA with 2 mM Bru for the last 15 min in the presence of camptothecin. The sequencing reads from the isolated nascent Bru-containing RNA mapped throughout genes covering both exons and introns and a relative rate of transcription could be determined for all genes by integrating the number of reads throughout the gene dividing it by the length of the gene. The hit density was expressed as "reads per thousand base pairs per million reads" (RPKM) and represents the relative distribution of reads for a particular sample. When comparing the read distribution between the camptothecin-treated and the control sample, we observed that 1142 genes showed a more than 2-fold decreased relative rate of transcription while 919 genes showed a more than 2-fold increased relative transcription rate (Fig. 5.1A and

B). Whether the genes found to have increased relative rates of transcription are truly synthesizing at an absolute higher rate is not clear since the data generated from Bru-Seq represents the distribution of reads rather than absolute expression values. Therefore, when synthesis is reduced in the body of large genes, sequencing reads must accumulate elsewhere (in small genes and at the beginning of large genes).

The data show an obvious negative correlation between read intensity and gene size following camptothecin treatment. The average size of the genes with more than a 2-fold decreased relative transcription rates was 136,355 bp. (Fig. 5.1C). However, there was also a subset of smaller genes showing reduced expression after camptothecin treatment with some histone genes and genes involved in the mitotic phase of the cell cycle such as CCNB1, CDK1, AURKA and AURKB highly enriched in this group (Fig. 5.1D). The average genomic size of the 919 genes showing increased relative transcription rates following camptothecin treatment was 8,927 bp. These findings are consistent with a mechanism of action for camptothecin as an inhibitor of transcription elongation without inhibiting transcription initiation.

5.4.2 Camptothecin affected expression of ncRNA and enhancer RNA (eRNA), transcription termination and splicing

In addition to inhibiting the elongation of protein-coding genes, camptothecin inhibited transcription elongation of primary microRNA transcripts and long non-coding RNAs (lncRNAs). For many short genes where no inhibition of elongation was apparent, transcription read-through past the 3' poly(A) site was prominent (Fig. 5.2A) suggesting a role of topoisomerase I in transcription termination in these genes (Durand-Dubief et al., 2011). Some genes showed a more pronounced splicing activity during the labeling period in the presence of camptothecin compared to untreated cells (Fig. 5.2B). It is possible that the reduced rate of elongation in the

presence of camptothecin allows more time for the splice junctions to be identified and spliced in a co-transcriptional manner (Listerman et al., 2006). However, the apparent enhancing effect of camptothecin on splicing was not observed for all genes.

Many genes in mammalian cells have been shown to generate divergent promoter upstream transcripts (PROMPTs) (Preker et al., 2008; Paulsen et al., 2013b). The expression of some PROMPTs was dramatically enhanced by camptothecin treatment (Fig. 5.2C). Furthermore, many divergently transcribed genes showed coordinate enhancement of initiation, suggesting that the negative superhelicity expected to accumulate in the wake of transcription in the absence of topoisomerase I activity enhances transcription initiation. Finally, many known and putative enhancer elements, such as the 5' FOS enhancer, showed increased generation of enhancer RNA (eRNA) in the presence of camptothecin (Fig. 5.2D). The functional consequence of the enhanced generation of eRNA following camptothecin treatment is not clear since the relative transcription rate of the FOS gene was not induced despite increased eRNA generated.

5.4.3 Transcription recovers as a wave from the 5' end following camptothecin removal

The trapping of topoisomerase I on DNA by camptothecin is thought to be a partially reversible event (Pommier, 2006). To explore whether the removal of camptothecin reverses its effects on transcription, we used Bru-Seq to examine the nascent RNA transcriptome in cells following drug washout. To get an aggregate picture of the effect of camptothecin on nascent RNA synthesis of multiple genes, we selected genes larger than 100 kb that generate a signal $>$ 1 RPKM. We aligned the Bru-Seq data from the transcriptional start site (TSS) of these genes and observed that bromouridine-labeling in the presence of camptothecin generated enhancement of

reads in the first 10 kb of the genes while the signal further downstream in the genes was severely suppressed (Fig. 5.3A). These results suggest that the inhibition and trapping of topoisomerase I by camptothecin does not inhibit initiation of transcription but severely inhibits elongation. When camptothecin was washed out and cells were labeled with bromouridine for 15 min in the absence of the drug, transcription reads spread from the 5'-end into the gene while no recovery of signal was observed further downstream in the gene. Following washout of the drug and incubation for 15 minutes in drug-free media and then labeling nascent RNA for the following 15 minutes, the transcription wave moved further into the gene in a 5'-3' direction. Again, no recovery of signal was observed further downstream into the gene. Interestingly, the transcription wave spreading from the 5'-end and into the body of the genes had an elongation rate of approximately 1.1-1.3 kb/min (Fig. 5.3). This is a slower rate than the estimated elongation rate of around 2 kb/min in cells under normal conditions (Danko et al., 2013). If elongating RNA polymerases collide with trapped topoisomerases, irreversible DNA damage may be induced that would require further processing (Wu and Liu, 1997). It is possible that this reduced rate of elongation observed following camptothecin treatment and washout is due to the requirement of repair of Top1/camptothecin-induced DNA damage before resumption of elongation can take place.

5.4.4 No apparent defect in the recovery of RNA synthesis in CS-B cells following camptothecin reversal

It has been shown that cells derived from Cockayne syndrome patients are hypersensitive to camptothecin [17]. This hypersensitivity is linked to an enhanced induction of DSBs in S-phase as replication forks "collide" with trapped Top1 complexes (Squires et al., 1993). Some studies have also shown that the recovery of

RNA synthesis is slower in CS cells (Squires et al., 1993; Horibata et al., 2011; Lin et al., 2013) while other studies have found no defect in RNA synthesis recovery in CSB-deficient cells (Sakai et al., 2012). Using Bru-Seq we here tested whether the recovery of nascent RNA synthesis in CS-B fibroblast cells following camptothecin treatment and reversal differed from the recovery in normal human fibroblasts. Analysis of the aggregate transcription signal of genes at least 100 kb or longer showed that CS-B cells recovered RNA synthesis in a wave from the 5'-end of these genes in a similar fashion as the normal fibroblasts (Fig. 5.4A). This was also apparent for the individual genes SMAD3, TLE4, POL1, CD44 and MEIS1 (Fig. 5.4). In addition, no recovery of transcription occurred from the bodies of the genes suggesting that initially blocked RNA polymerases are not able to resume elongation following camptothecin removal. The apparent normal recovery of RNA synthesis in these CS-B cells were in sharp contrast to the defective recovery of nascent RNA synthesis in these cells following UV-irradiation (unpublished data).

5.4.5 Camptothecin affected cancer-relevant gene expression

Performing DAVID gene enrichment analysis we found that camptothecin-induced genes coding for elements of the ribosome, mitochondrion and the p53 and apoptosis signaling pathways were highly represented (Figure 5.5). The set of genes found to be inhibited shortly after camptothecin treatment was enriched for phosphoproteins, proto-oncogenes, and genes involved in the mitotic cell cycle, ubiquitin conjugation and anti-apoptosis. Some representative large proto-oncogenes inhibited by camptothecin are shown in Figure 5.6A. It has been shown that blockage of transcription elongation by camptothecin triggers a stress response leading to the rapid accumulation of p53 accompanied by phosphorylation of the Ser15 site and acetylation of the Lys382 site (Ljungman et al., 2001). In support of camptothecin inducing

a p53 response in human fibroblasts, we found that camptothecin induced genes in the p53 signaling pathway, including CDKN1A (p21), MDM2, BTG2 and FAS (Figure 5.6B). Some of these genes were induced already during the camptothecin treatment while some genes, like CDKN1A and MDM2, showed induced expression only following reversal of drug treatment. Camptothecin reduced the relative transcription rates of large anti-apoptotic genes and enhanced expression of a set of smaller sized pro-apoptotic genes (Figure 5.6C). Pro-apoptotic genes are generally more compact compared to anti-apoptotic genes (McKay et al., 2004), thus agents preferentially reducing expression of large genes by blocking transcription elongation are expected to shift the balance of gene expression in favor of apoptosis. Similar patterns of increased and decreased relative gene expression following camptothecin treatment and reversal were found for CS-B cells. There were, however, some differences between normal human fibroblasts and the CS-B cells were observed such as a lack of reduced GLI2 expression and no induction of the p53-regulated genes DUSP5, FAS, MDM2 and TRIM22 in CS-B cells.

5.5 Discussion

Camptothecin and its derivatives are FDA approved anti-cancer drugs used to treat a variety of tumors (Pommier, 2013). They act by trapping topoisomerase I complexes on DNA rather than inhibiting enzymatic function, since RNAi knockdown of Top1 does not reduce cell survival to the same degree as camptothecin treatment (Nitiss and Wang, 1988; Pommier, 2013). In this study, we used Bru-Seq to explore the acute effects of camptothecin on various aspects of transcription and found that camptothecin (i) inhibited elongation of transcription, (ii) stimulated transcriptional read-through past the 3-end of small genes, (iii) enhanced expres-

sion of eRNA from certain enhancer elements (iv) induced the p53 response and (v) shifted the balance of expression of apoptosis-regulatory genes in favor of apoptosis. Importantly, transcription recovered with a reduced elongation rate as a wave from the 5-end of the gene with no apparent recovery of synthesis from RNA polymerases blocked in the body of the genes. We found no evidence that the recovery of RNA synthesis was different in CS-B fibroblasts which is in sharp contrast to the recovery of RNA synthesis in these cells after UV light (unpublished data). Thus, the mechanisms responsible for the recovery of RNA synthesis following camptothecin removal are fundamentally different from those required following UV-irradiation suggesting that transcription-coupled repair has no major role in the restart of transcription following camptothecin removal. It is therefore conceivable that the observed hypersensitivity of CS-B cells to camptothecin is related to some role of the CSB protein during recovery of replication rather that in the recovery of transcription (Squires et al., 1993).

The inability of cells to restart transcription from within the body of genes suggests that blocked RNA polymerases are discarded rather than recycled. This will preferentially set back the expression of large genes even following a limited exposure of cells to camptothecin. Interestingly, many proto-oncogenes and anti-apoptotic genes belong to the class of genes preferentially inhibited by camptothecin. The model that emerges is that poisoning of Top1 by camptothecin results in the inhibition of large proto-oncogenes, enhanced expression of small pro-apoptotic genes and activation of the p53 pathway (Figure 5.6D). Knowledge of the size of the oncogenes that drive carcinogenesis and are important for survival of cancer cells in a given tumor may be used to select patients who would specifically benefit from camptothecin treatment and to rationally combine camptothecin with other treatment modalities.

For example, the expression of the large BRCA1-associated RING domain protein 1 gene (BARD1) was reduced by camptothecin. Such suppression would be expected to suppress homologous recombination and thus should lead to increased susceptibility to PARP inhibitors or radiation therapy. Indeed, it has been shown that combining camptothecin with PARP inhibitors or radiotherapy improves tumor control (Chen et al., 1999; Bowman et al., 2001; Zhang et al., 2011; Pommier, 2013).

5.6 Acknowledgments

We thank Manhong Dai and Fan Meng for administration and maintenance of the University of Michigan Molecular and Behavioral Neuroscience Institute (MBNI) computing cluster and the personnel at the University of Michigan Sequencing Core for technical assistance.

5.7 Grant Support

This work was supported by funds from National Institute of Environmental Sciences (1R21ES020946) and National Human Genome Research Institute (1R01HG006786).

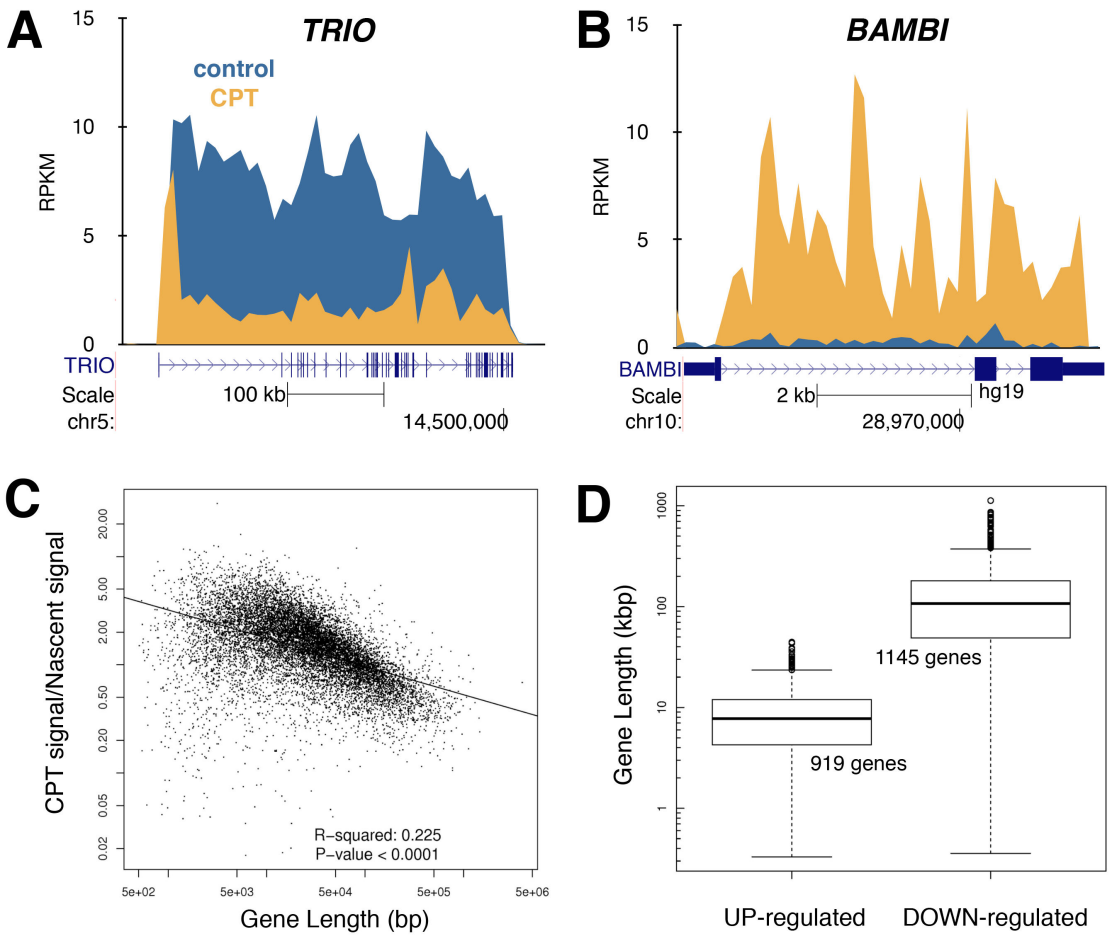


Figure 5.1: Gene size is a major contributing factor to the effects of camptothecin on RNA synthesis. Human fibroblasts were treated with $20\mu M$ camptothecin for 45 min with 2 mM Bru added during the last 15 min of camptothecin treatment to label nascent RNA followed by Bru-Seq. (A), Long genes, such as TRIO, exhibit elongation defects, but not transcription initiation, after camptothecin treatment. (B), Short genes, such as BAMBI, show a relative increase of RNA synthesis following camptothecin treatment. (C), Effect of camptothecin on relative transcription as a function of gene size. Ratio of Bru-Seq signal of individual genes in camptothecin-treated over control cells as a function of gene size. Longer genes are inhibited preferentially over shorter genes. (D), The median length of genes induced $>$ 2-fold by camptothecin (919 genes) is 8,927 bp, whereas genes down-regulated $>$ 2-fold (1,145 genes) have a median length of 136,355 bp. The gene maps are from RefSeq Genes (UCSC genome browser)

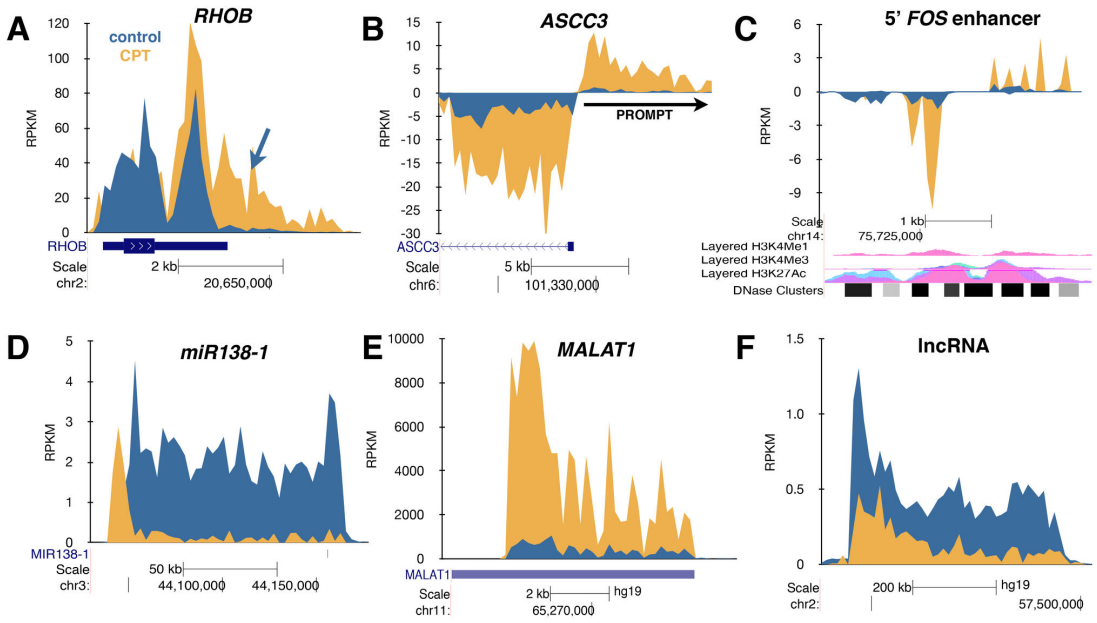


Figure 5.2: Effect of camptothecin on transcriptional readthrough and synthesis of PROMPTs and eRNA. As in Figure 5.1, human fibroblasts were treated with $20\mu M$ camptothecin for 45 min with 2 mM Bru added during the last 15 min of camptothecin treatment to label nascent RNA followed by Bru-Seq. (A), Transcriptional readthrough of the termination site of the RHOB gene induced by camptothecin. (B), Enhanced initiation of the ASCC3 gene and coincident upregulation of divergent upstream PROMPT RNA. (C), Enhanced expression of eRNA from the 5'-upstream enhancer of FOS by camptothecin. (D), Camptothecin inhibits the transcription of the primary transcript of miRNA138-1. (E), Camptothecin induces transcription of the ncRNA MALAT1. (F), Camptothecin inhibits the transcription of a very long unannotated ncRNA on chromosome 2. The gene maps are from RefSeq Genes (UCSC genome browser).

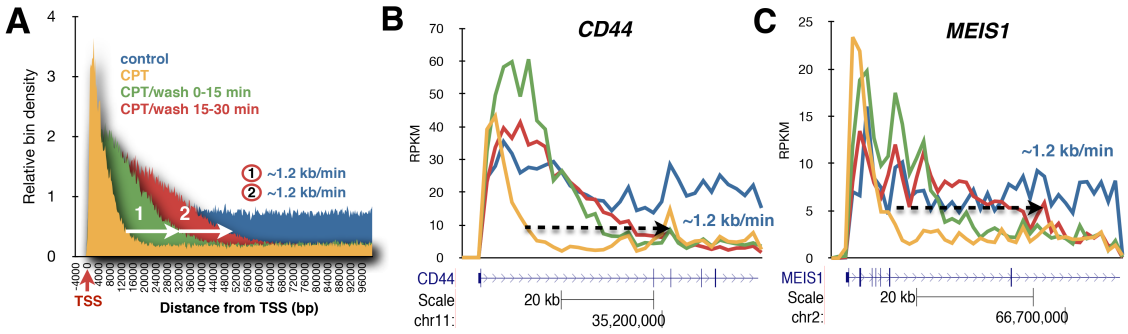


Figure 5.3: Effect of camptothecin reversal on RNA synthesis. (A), Aggregate view of RNA synthesis of genes larger than 100 kb in normal human fibroblasts with the genes aligned by transcriptional start sites (TSS). RNA synthesis recovers as a wave in a 5-to-3 direction following camptothecin removal with no apparent recovery of RNA polymerases stalled in the body of the genes. Elongation rates of the recovering transcription wave was estimated to be 1.2 kb/min. (B), Wave of recovery of RNA synthesis can be seen advancing from the 5-end of the CD44 gene with no apparent recovery in the body of the gene. The front of the transcription wave extended some 35 kb during the first 30 min recovery resulting in an elongation rate of about 1.2 kb/min. (C) Similar elongation rates after camptothecin removal were found for the MEISE1 gene. Color key: Blue, control (30 min Bru labeling); Yellow, Bru labeling during the last 15 min of a 45 min camptothecin treatment; Green, 45 min camptothecin treatment followed by a drug washout and 15 min of Bru labeling; Red, 45 min camptothecin treatment followed by a drug washout, 15 min incubation, and finally 15 min Bru labeling.

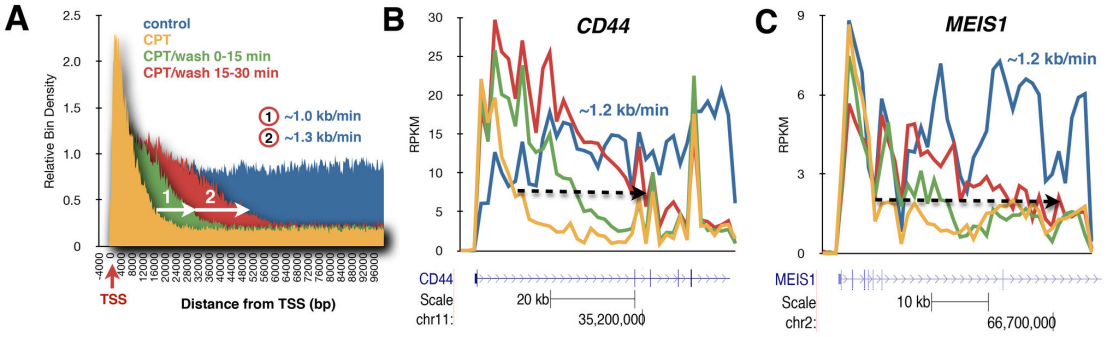


Figure 5.4: Effect of camptothecin reversal on RNA synthesis in Cockayne syndrome cells. (A), Aggregate view of RNA synthesis of genes larger than 100 kb in CS-B cells with the genes lined up by transcriptional start sites (TSS) as in Figure 3. Elongation rates of the recovering transcription wave was estimated to be 1.0-1.3 kb/min. Individual genes in fibroblasts from a CS-B individual showing similar recovery rates as in fibroblasts from a normal individual for (B), CD44 and (C) MEIS1. Color key as in Figure 5.3.

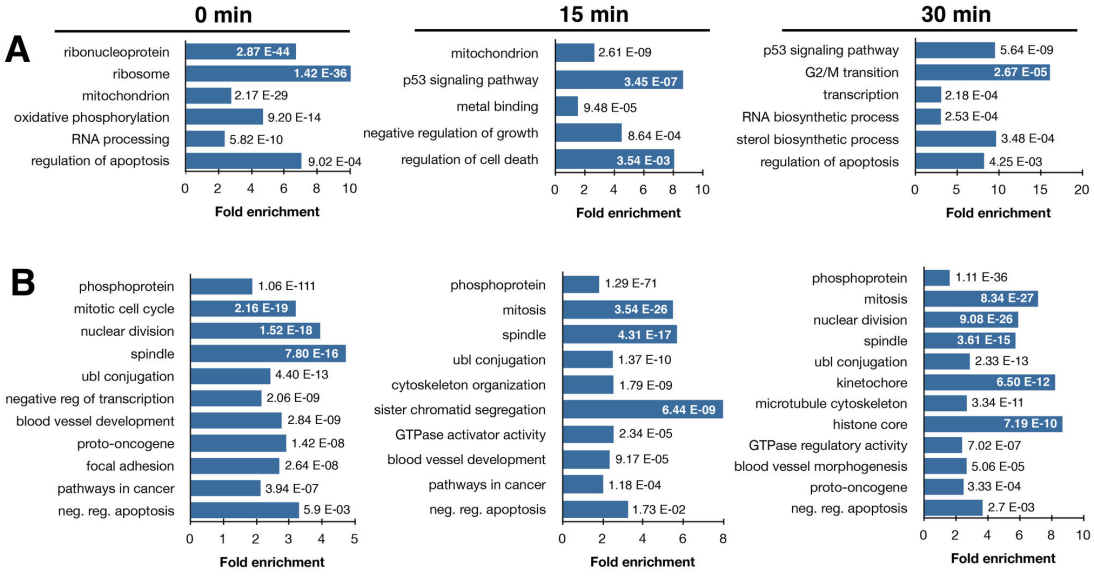


Figure 5.5: Pathway enrichment for genes following camptothecin treatment and reversal. (A) Pathways represented by genes up-regulated at least 2-fold, and (B) down-regulated at least 2-fold following camptothecin treatment and recovery. Human fibroblasts were treated with $20\mu M$ camptothecin for 45 minutes and incubated for the last 15 min with 2 mM Bru ("0 min"), incubated for 15 min with Bru following the removal of camptothecin ("5 min") or incubated for 15 min with Bru following a 45 min treatment, a wash and a 15 min recovery ("30 min"). Enrichment analysis was performed using DAVID (`david.abcc.ncifcrf.gov`) and the numbers shown represents the p-values for enrichment.

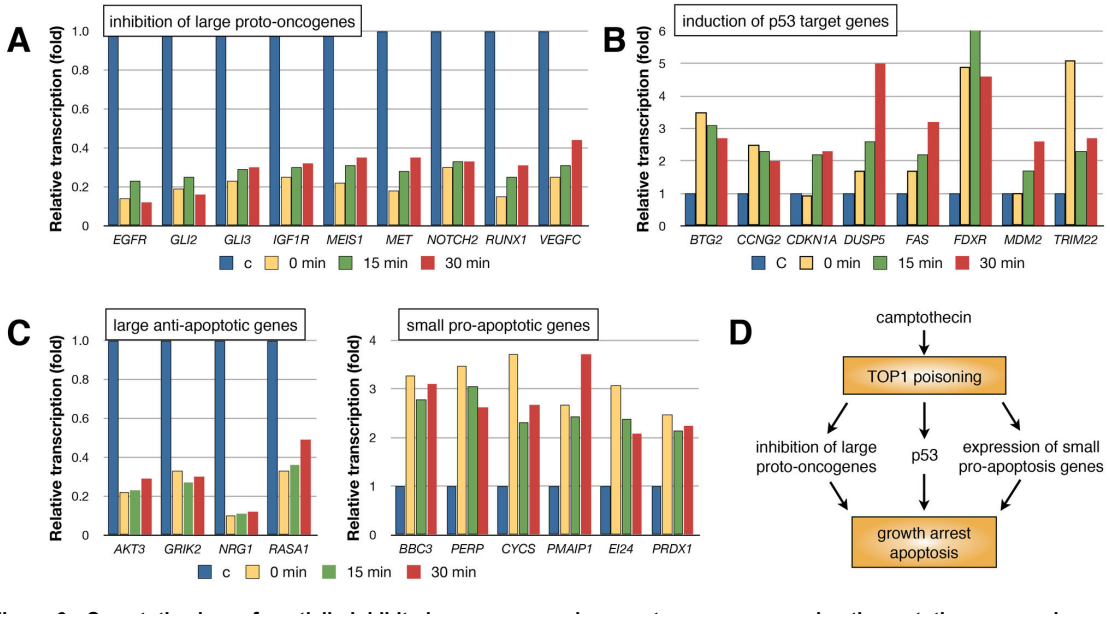


Figure 5.6: Camptothecin preferentially inhibits large genes such as proto-oncogenes and anti-apoptotic genes, enhances the relative expression of small pro-apoptotic genes and activates the p53 response. (A), Examples of large proto-oncogenes inhibited by camptothecin and showing no recovery (or slow recovery) following drug removal. (B), Examples of p53 target genes induced following camptothecin treatment. (C), Examples of large anti-apoptotic genes showing reduced relative transcription (left) and examples of small pro-apoptotic genes showing enhanced relative transcription following camptothecin treatment (right). (D), Model of mechanisms by which camptothecin may induce cell death or inhibit cell growth. Camptothecin triggers a p53 transcriptional response and selectively inhibits large proto-oncogenes and survival genes. The data is color coded where blue represents control (C), yellow represents 15 min Bru-labeling at the end of a 45 min camptothecin treatment with no recovery ("0 min"), green represents drug washout and 15 min Bru-labeling immediately after washout ("15 min") and finally red represents labeling 15-30 minutes following washout ("30 min").

CHAPTER VI

Rate of transcriptional elongation associates with H3K79me2 and H4K20m1 epigenetic marks

6.1 Abstract

The rate of transcription elongation plays an important role in the timing of expression of full-length transcripts as well as in the regulation of alternative splicing. In this study we coupled Bru-seq technology with 5,6-dichlorobenzimidazole 1-β-D-ribofuranoside (DRB) to estimate the elongation rates of over 2,000 individual genes in human cells. This technique, BruDRB-seq, revealed gene-specific differences in elongation rates with a median rate of around 1.5 kb/min. We found that genes with rapid elongation rates showed higher densities of H3K79me2 and H4K20me1 histone marks compared to slower elongating genes. Furthermore, high elongation rates had a positive correlation with gene length, low complexity DNA sequence and distance from nearest active transcription unit. Features that negatively correlated with elongation rate included the density of exons, long terminal repeats, GC content of the gene and DNA methylation density in the bodies of genes. Our results suggest that some static gene features influence transcription elongation rates and that cells may alter elongation rates by epigenetic regulation. The BruDRB-seq technique

Official citation:
Veloso, A., Kirkconnell, K., Magnuson, B., Biewen, B., Paulsen, M.T., Wilson, T.E., Ljungman, M, Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Research, In Press*

offers new opportunities to interrogate mechanisms of regulation of transcription elongation.

6.2 Introduction

Gene transcription in eukaryotes is the highly regulated process by which RNA polymerase II (RNAPII) uses DNA as a template to produce RNA. The stages of transcription include initiation, elongation, and termination, the control of which influences gene expression. Mechanisms of transcription initiation have been studied in detail and much is known about transcription factor activation and binding, pre-initiation complex formation, and RNAPII recruitment (Shandilya and Roberts, 2012). Furthermore, the critical roles of regulatory sequences such as enhancer elements for developmental and tissue-specific gene regulation (Spitz and Furlong, 2012) and the three dimensional organization of the transcription machinery have been characterized to some level (Sutherland and Bickmore, 2009). However, the importance of regulation of the rate of transcription elongation is poorly understood.

Activation of specific gene programs, such as those regulating early organism development, is thought to depend on gene size to accomplish a temporal expression pattern after simultaneous transcriptional activation (Swinburne and Silver, 2008). A proposed mechanism to delay the generation of mature RNA is the inclusion of introns of various sizes (Seoighe and Korir, 2011; Takashima et al., 2011). To fine-tune this timing mechanism in gene expression, cells may adjust the rates of transcription elongation, splicing, nuclear export, and ribosome access. The rate of transcriptional elongation has also been tied to alternative splicing patterns, where high transcription elongation rates favor exclusion of alternative exons while slow elongation rates correlate with their inclusion (Close et al., 2012; Shukla and Oberdoerffer, 2012).

Previous studies have measured in vivo RNAPII elongation rates in mammals using a variety of techniques including RT-PCR (Singh and Padgett, 2009), tiling microarrays (Wada et al., 2009), and fluorescent labeling (Darzacq et al., 2007b). These studies have been limited to a single or a small number of genes and have reported a wide range of elongation rates. A recent study utilized GRO-seq to assess elongation rates of a much larger set of genes activated by estradiol or tumor necrosis factor (TNF) and demonstrated a broad range of transcriptional elongation rates among the set of activated genes, supporting the notion that elongation rates may be regulated (Danko et al., 2013).

Here we utilize BruDRB-seq to assess transcription elongation rates genome- wide. This technique involves the transient inhibition of initiated RNAPII prior to elongation using 5,6-dichlorobenzimidazole1-β-D-ribofuranoside (DRB) (Singh and Padgett, 2009). Following drug removal, RNA polymerases enter the elongation phase in a synchronized manner and nascent RNA is labeled with bromouridine (Bru), isolated with anti-BrdU antibodies and subjected to deep sequencing. By measuring the width of the transcription "wave" generated during the labeling period, the transcription elongation rates of all expressed genes longer than 40 kb were assessed. Our study provides the largest data set so far reported of genome-wide elongation rates in multiple cell lines. We found that high transcription elongation rates correlated with specific gene features as well as with histone modifications such as di-methylation of lysine 79 of histone H3 (H3K79me2) and mono-methylation of lysine 20 of histone H4 (H4K20me1). These results indicate that cells may be able to fine tune transcription elongation rates by epigenetic regulation.

6.3 Results

6.3.1 Measuring elongation rates globally reveals variation among genes

To study the elongation rates of RNAPII genome-wide, we developed BruDRB-seq, a technique based on nascent RNA labeling with Bru and assaying by deep sequencing (Paulsen et al., 2013a,b). Following a 60 minute treatment of the cultured cells with DRB to arrest RNAPII at promoter-proximal sites, the drug was washed out and the cells were incubated with Bru for 10 minutes either directly or after a 10 min recovery period. Cells were lysed in TRIzol and total RNA was isolated followed by specific capturing of Bru-labeled RNA using anti-BrdU antibodies conjugated to magnetic beads. The captured Bru-labeled RNA was then reverse- transcribed and the resulting cDNA library was subjected to deep sequencing using the Illumina HiSeq 2000 platform.

In Figure 6.1A (control), all expressed genes of at least 50 kb in length in the diploid human fibroblast cell line HF1 are represented by median normalized expression (an aggregate view). As expected for nascent RNA, the signal was fairly evenly distributed throughout the first 50 kb of these genes. Following a 60 minute DRB treatment with Bru labeling during the last 10 minutes of treatment, a substantially lower yield of reads was obtained indicating that transcription was severely reduced (Fig. 6.1A, 0 min). Following drug removal, a synchronized wave of transcription was observed moving out from the promoter (Fig. 6.1A, 10 min) and this wave moved further during the next 10 min (Fig. 6.1A, 20 min). These results demonstrate both the reversibility of DRB and synchronicity of transcription recovery.

In order to measure elongation rates of individual genes in a genome-wide fashion, we used an inference method based on a three state Hidden Markov Model (HMM) (Day et al., 2007; Danko et al., 2013). The HMM was designed to identify three dis-

tinct regions of each gene: (A) the region immediately upstream of the transcription start site (TSS), (B) the advancing wave, and (C) the region of low transcription downstream of the advancing transcription wave (Fig. 6.1B). Quantile normalization was performed on the BruDRB-seq trace prior to HMM analysis to eliminate any effect of a gene's expression level on the analysis (see Methods). After applying the HMM analysis to the data from the 10 min wave in HF1 cells, we ordered the genes according to their calculated elongation rates and found them to be quite variable (Fig. 6.1C). Examples of transcription waves moving from the promoters into the bodies of three individual genes following DRB removal are shown in Figure 6.1D-F.

DRB inhibits the transition of RNAPII from the initiation/promoter paused stage into the elongation phase by blocking the phosphorylation of the C-terminal domain (CTD) of RNAPII (Dubois et al., 1994). However, DRB does not inhibit elongating RNA polymerases and thus, DRB treatment results in a time-dependent clearing out of transcription from the promoter with a receding wave of unaffected actively transcribing polymerases. For transcribed genes longer than 200 kb in HF1 cells treated with DRB for 60 min, the receding transcription wave can be clearly observed (Fig. 6.2A, yellow). The elongation patterns of two large genes expressed in HF1 cells, MYO1B and TLE4, are shown in Figure 6.2B and 6.2C, respectively.

By incorporating a fourth state in the Hidden Markov Model representing the receding wave of transcription, we were able to analyze the correlation between the advancing and the receding waves. A visual comparison between the states predicted by the HMM (Fig. 6.2D) and the normalized signal observed in those genes (Fig. 6.2E) indicates that predictions of the model were reasonably accurate. We compared elongation rates calculated both via the advancing (state "B") and the receding (state "D") wave and found that they correlate, albeit with high variability

(Fig. 6.2F). Because the trailing edge of the receding wave is not as well defined as the advancing wave and because we can include more genes by leaving the receding wave out of the HMM, we decided to focus on the advancing wave (3-state HMM predictions) as our metric for elongation.

6.3.2 Elongation rates are similar in different cell lines

Transcription elongation rates have been explored in a limited number of genes and only in a few cell lines Ardehali et al. (2009); Singh and Padgett (2009); Danko et al. (2013). In this study, we used five cell lines and BruDRB-seq to assess transcription elongation rates genome-wide. Three of these cell lines are human fibroblasts and two cell lines, K562 and MCF-7, are cancer-derived. HF1 and TM cells are normal human fibroblasts while Cockayne syndrome B cells (CS-B) have a genetic defect in the ERCC6 gene, which encodes the CSB protein, resulting in a defect in transcription-coupled DNA repair. It has been suggested that the CSB protein associates with the elongation transcription complex and in vitro results suggest that CSB enhances the rate of RNAPII elongation (Selby and Sancar, 1997). The median elongation rate was found to be similar across the five cell lines, with HF1, CS-B, and K562 being nearly identical (1.25 kb/min), and TM and MCF-7 rates slightly higher than the other cell lines (1.75 kb/min) (Fig. 6.3A). In addition, there was a positive correlation of elongation rates of individual genes between the cell lines when performing a pair-wise comparison (Fig. 6.3B). As examples, similar transcription elongation rates for the ACTN4 and PTEN genes across the five cell lines are shown in Figure 6.3C.

A clustering method was used to identify similarities in the observed elongation rates across the five cell lines. Quantile-normalized elongation rates of approximately 800 genes expressed in all five cell lines were put into a k-medoids algorithm to cluster these genes into three groups based on similarities in elongation rates. The

gene groups selected by the algorithm were clearly distinguished by their overall elongation rates, with a fast, a slow and a variable intermediate group (Fig. 6.3D). These observations indicate that elongation rates of individual genes are considerably conserved among cell lines.

6.3.3 Gene Set Enrichment

To determine whether genes with similar functions or belonging to a particular pathway have similar transcription elongation rates, we assessed gene enrichment clustering among genes with similar elongation rates. We centered the average elongation rate along its own mean and provided these values and the gene symbols as pre-ranked lists to the Gene Set Enrichment Analysis (GSEA) tool (Subramanian et al., 2005). We searched for enrichment in positional, curated (BioCarta and KEGG), gene ontology, and oncology signature gene sets obtained from the Molecular Signatures Database. We used the permissive false discovery rate (FDR) p-value suggested by the authors of GSEA ($p <= 25\%$) (Subramanian et al., 2005), and focused on the gene sets that were enriched in at least three of the cell lines. We found that genes related to organic acid and carboxylic acid metabolism were enriched among the genes with slow elongation rates. Furthermore, genes related to regulation of actin cytoskeleton and to leukocyte trans-endothelial migration were enriched among genes with higher elongation rates.

6.3.4 Gene sequence features correlate to elongation rates

Since at least half of the genes used for the clustering analysis described in Figure 6.3D grouped strongly by elongation rate independent of cell type, we reasoned that the rate of elongation for some genes may be correlated with specific DNA sequence features of the transcribed DNA. Certain features, such as splice site sequences and

sequences with a propensity to form G-quadruplexes, have been implicated to affect RNAPII elongation rates in vitro (Belotserkovskii et al., 2013). We investigated the correlation between elongation rates and several sequence features, including GC content, exon density, regions of repetitive DNA, and sequences computationally predicted to form non-B DNA structures. For each gene, the density of each feature from the TSS to the end of the advancing wave was calculated and compared to its elongation rate. Because several of these features were not randomly distributed throughout the first 40 kb of the genes, a simple regression analysis was inadequate to establish a correlation. To confirm that the correlations found with these features were not solely due to the non-random distribution of the feature, we conducted a permutation analysis using a FDR-corrected p-value of 0.05 as the threshold for significance.

In all cell lines analyzed, exon density (and therefore splice site density) was negatively correlated with elongation rate (Table 6.1). GC content and the density of long terminal repeat sequences were also negatively correlated with elongation rates in at least 3 cell lines. It is possible that a higher GC content reduces elongating rates due to a higher energy requirement for breaking three hydrogen bonds between G and C versus two for A and T. The only DNA sequence feature to positively correlate with elongation rate was a high density of low complexity sequences (stretches of mono- or di-nucleotide repeats).

6.3.5 Role of gene neighborhoods and genomic organization

Gene expression is often determined by whether the gene is located in an open (euchromatic) or condensed (heterochromatic) chromatin configuration. We asked whether transcription elongation rates are influenced by proximity to nearby genes, chromosomal regions, or three-dimensional organization, which we collectively refer

to as "gene neighborhoods". To examine the effects of gene neighborhoods on elongation, we compared the measured elongation rates of genes and their proximity to neighboring genes on either strand both upstream and downstream of the gene (Table 6.1). We found a positive correlation between elongation rate and the distance to other genes. Thus, active transcription nearby has a negative impact on elongation rate. Interestingly, we also found that gene length is positively correlated with elongation rate though it is unclear how longer genes are identified or marked for faster transcription. Next we examined whether there was a correlation between the elongation rates of neighboring genes and found no statistically significant relationship between the elongation rates of neighboring transcribed genes (Supplemental Fig. S6.1A). Furthermore, inspection of the distribution of genes and their associated elongation rates along the different chromosomes suggests that genes with high or low transcription elongation rates were distributed randomly throughout out the genome (Supplemental Fig. S6.1B).

The above analyses addressed the influence of gene proximity as defined by a linear chromosome, but does not consider the three dimensional organization of genes within the nucleus. Genes that are linearly distant, even on completely different chromosomes, may interact due to long distance DNA looping and may be transcribed by the same transcription machinery (Dekker et al., 2013). To assess whether genes associated with each other in the same transcriptional factory have similar elongation rates, we used publicly available data from chromatin interaction analysis by pair-end tag sequencing (ChIA-PET) for K562 and MCF-7 cells. The elongation rates of genes that were shown by ChIA-PET to co-localize to the same transcription machinery were plotted against each other as gene 1 vs. gene 2 (Supplemental Fig. S6.1C). The results show that there was no significant correlation between elongation rate and

chromatin interactions.

6.3.6 Elongation rates are related to epigenetic modifications

While we found that some DNA sequence features correlated with elongation rate (Table 6.1), some genes showed varied elongation rates across cells lines (Fig. 6.3D). Thus, the elongation rates of some genes may be regulated in a cell-type specific way and we hypothesized that cells may regulate transcription elongation rates by specific epigenetic modifications. We first explored whether the level of DNA methylation in the body of genes correlated with transcription elongation rates. While methylation of CpG islands in promoter regions has been implicated in gene silencing, the function of CpG methylation in the body of genes is poorly understood (Jones, 2012). We compared elongation rates obtained with BruDRB-seq to published genome-wide CpG methylation patterns (bisulfite sequencing) data for K562 and MCF-7 cells (ENCODE) and found that genes with high levels of DNA methylation tended to elongate at slower rates. This effect, however, was only noticeable when analyzing CpG sites with high occurrence of methylation (at least 90%), as the correlation was not significant when including sites with lower occurrence (e.g. at least 50% methylation) in the analysis.

To explore whether fast or slow transcription elongation rates may associate with the presence of specific histone modifications, we divided the genes into four quartiles according to elongation rates and compared them with ChIP- seq data available through ENCODE for different histone marks that have been implicated in transcription regulation (Ernst et al., 2011; ENCODE Project Consortium et al., 2012). Of the histone marks tested, only H3K79me2 and H4K20me1 were found to show a significant positive correlation with elongation rates (Fig. 6.4A&B; Supplemental Fig. S6.2). These histone marks have been shown to be linked to active transcription

(Rao et al., 2005; Smolle and Workman, 2013) but they have not previously been shown to influence the rate of elongation. We did not observe a significant correlation between transcription elongation rates and the densities of tri-methylation of lysine 36 of H3 (H3K36me3) (Fig. 6.4C) or RNAPII (Fig. 6.4D). Since H3K36me3 and RNAPII densities are known to correlate with levels of gene expression, our data suggest that high elongation rates are not merely reflecting high expression levels although a weak positive correlation was observed between transcript output measured with Bru-seq and transcript elongation rate (Kendall's Tau=0.23, Supplemental Fig. S6.3). We found no significant relationship between the density of other common histone marks or transcription factors and transcription elongation rate (Supplemental Fig. S6.2).

6.4 Discussion

Transcription elongation has recently drawn attention due to its potential role in the timing of gene expression and the regulation of alternative splicing (Mason and Struhl, 2005; Darzacq et al., 2007b; Singh and Padgett, 2009; Wada et al., 2009; Danko et al., 2013). Here we describe a novel technique, BruDRB-seq, to measure RNAPII elongation rates genome-wide. BruDRB-seq is based on DRB-induced arrest of RNAPII at promoter sites (Singh and Padgett, 2009) followed by synchronized release after drug removal. In five different human cell lines, median elongation rate estimationsranged from 1.25-1.75 kb/min. These transcription elongation rates, which were estimated from over 2000 genes in each cell line, are somewhat lower than previously estimated in human cells (Singh and Padgett, 2009; Danko et al., 2013). It is possible that the genes analyzed by Danko et al., which had been induced by estradiol or TNF, showed a higher elongation rate due to being in an induced state

where higher transcription and RNAPII densities promoted enhanced elongation rates. Our results also differ from a previously published study that implicated a role of the CSB protein in transcription elongation (Selby and Sancar, 1997). In our study, CS-B fibroblasts did not show a distinctly different elongation rate than the other cell types suggesting that the CSB protein is not generally required for promoting rapid transcription elongation in cells, though it may be an important regulator of elongation for select genes.

Our BruDRB-seq data indicate that there is a broad range of transcription elongation rates in different genes in human cells (Fig. 6.3). However, the elongation rates for individual genes were reasonably conserved across the different cell lines. We speculated that this conservation may be driven by specific physical features, such as DNA sequence, gene length or genomic position. Among the genetic features linked to elongation rate was exon density, which correlates with slow elongation. This supports the idea that RNAPII slows down at splice site junctions, which would promote exon definition and alternative splicing (Shukla and Oberdoerffer, 2012). Although it has been shown that non-B DNA sequences can have a negative impact on transcription in vitro (Belotserkovskii et al., 2013), we did not find a correlation between elongation rate and potential non-B DNA sequences, though some of these sequences may not exhibit non- B DNA conformations in vivo. Furthermore, our results suggest that if a gene is located near another transcribing gene, its transcription elongation rate is decreased. It has been proposed that transcription of a downstream gene could lead to unwinding of DNA through the induction of negative supercoiling (Ljungman and Hanawalt, 1992, 1995) which may affect the elongation of a proximal gene. Conversely, if neighboring genes are simultaneously transcribed in a head to head fashion, DNA topological barriers could emerge. Indeed, inhi-

bition of DNA topoisomerase I, which relaxes torsional tension induced during the transcription process, has been shown to severely inhibit transcription elongation (Ljungman and Hanawalt, 1996; Veloso et al., 2013). Interestingly, suppression of the rate of elongation by nearby transcription was independent of the orientation of transcription. It is possible that the slower elongation rate in gene neighborhoods is due to competition for limiting pools of ribonucleotides rather than restraints caused by transcription-induced DNA supercoiling.

Our findings that the rates of transcription elongation of nearby genes or genes interacting via DNA looping did not correlate with each other suggest that elongation rates are primarily governed by gene-specific features and epigenetic modifications. We found that H3K79me2 and H4K20me1 were enriched in genes with higher elongation rates (Fig. 6.4), while the density of H3K36me3 marks, which have been implicated in transcription elongation (Guenther et al., 2007), did not correlate with elongation rate in our study. Furthermore, we did not observe a strong association between high elongation rates and high density of RNAPII (Fig. 6.4) or high levels of transcription in those genes (Supplemental Fig. 6.3). Thus, our data suggest that high elongation rates are not simply the result of high gene expression but rather, are governed by specific gene features and epigenetic modifications.

The histone mark H3K79me2 is regulated by the methylase DOT1L (Min et al., 2003), an epigenetic regulator implicated in somatic cell reprogramming (Onder et al., 2012). It was found that key genes involved in the induction of a mesenchymal cell state lost dimethylation of H3K79 without a change in their expression levels during this transition. It is possible that reducing the density of H3K79me2 marks results in a reduced elongation rate of these genes and that this is allowing these cells to transition into a new cell state. DOT1L has also been found to be associated with mixed

lineage leukemia (MLL) fusion proteins, resulting in aberrant methylation patterns of H3K79 and dysregulation of MLL targeted genes (Okada et al., 2005). Clinical trials are currently underway using DOT1L-targeting drugs in MLL (Anglin and Song, 2013). Methylation of H4K20 is regulated by the PR-SET7 methyltransferase in a cell cycle-dependent manner (Nishioka et al., 2002) and these modifications have been shown to play a role in DNA damage responses by attracting 53BP1 (Beck et al., 2012). Interestingly, both the H3K79me2 and H4K20me1 histone marks have been implicated in the regulation of replication origin firing (Tardat et al., 2010; Fu et al., 2013). Perhaps by regulating transcription elongation rates of long genes by H3K79 and H4K20 methylation, cells can fine-tune the firing of replication origins.

The size of a gene is a major determinant for how long it will take for transcription to be completed and for the gene to be expressed, and differences in gene lengths contribute to temporal expression patterns Swinburne and Silver (2008). Our study shows that transcription elongation rates are associated with the histone marks H3K79me2 and H4K20me1, providing a potential mechanism by which cells can fine- tune the temporal gene expression and alternative splicing patterns, despite fixed gene lengths. Future studies are needed to define the mechanisms by which cells regulate elongation rates through epigenetic modification and characterize pathological states whereby transcription elongation rates are dysregulated.

6.5 Methods

6.5.1 Cell culturing

HF1, hTERT immortalized foreskin-derived human fibroblasts, previously called NF (Paulsen et al., 2013a,b; Veloso et al., 2013), CS-B primary human skin fibroblasts (Coriell, GM00739) and TM, hTERT immortalized human skin fibroblasts (a gift from Dr. Tom Misteli, NCI) were grown in MEM supplemented with 10% FBS, L-

glutamine, vitamin mix and antibiotics. K562 human leukemia cells were grown in IMDM supplemented with 10% FBS and penicillin/streptomycin. MCF-7 human breast cancer cells were grown in high-glucose RPMI supplemented with 10% FBS.

6.5.2 Bru-seq and BruDRB-seq

The labeling of nascent RNA with bromouridine (Bru) was carried out as previously described (Paulsen et al., 2013a,b). The BruDRB-seq protocol differs from the Bru-seq protocol in that the drug 5,6-dichlorobenzimidazole 1-β-D- ribofuranoside (DRB, Sigma) is added to the media to a final concentration of $100\mu M$ and cells are incubated for 1 hour at 37°C. After the incubation with DRB, the cells were washed with PBS twice and nascent RNA was labeled in conditioned media containing 2 mM bromouridine (Bru) (Aldrich) for 10 min at 37°C. The cells were then directly lysed in TRIzol reagent (Invitrogen). K562 cells were grown in suspension so these cells were quickly spun down before being lysed in TRIzol. Total RNA was isolated and the Bru-labeled RNA was isolated from the total RNA by incubation with anti-BrdU antibodies (BD Biosciences) conjugated to magnetic Dynabeads (Invitrogen) under entle rotation for 1 hour at room temperature. Finally, cDNA libraries were made from the Bru-labeled RNA using the Illumina TruSeq library kit and sequenced using Illumina HiSeq sequencers at the University of Michigan DNA Sequencing Core. The sequencing and read mapping was carried out as previously described (Paulsen et al., 2013a,b).

6.5.3 Gene selection for elongation rate analysis

The Ensembl gene annotation (release 69) (Flicek et al., 2013) was used in this analysis and the annotation data was downloaded using the biomaRt package in the R environment (Durinck et al., 2005). All transcripts from genes with biotype

matching “protein coding”, “pseudogene”, “processed transcript” or “lincRNA” were initially selected to be used in the analysis. Transcripts were selected based on their length, and the transcript’s minimum acceptable length was 40 kb in the three state HMM analysis (see “Hidden Markov Model for elongation rate analysis section”) and 150 kb in the four state HMM analysis.

A potential source of error for the HMM analysis is the presence of additional TSSs either upstream or downstream of a given TSS. To address this issue, we first selected genes where the value 3’ of the TSS was at least 10 times higher than the value 5’ of the TSS. Second, we rejected genes that initiated transcription from an additional TSS within the analysis range (e.g. 40 kb in the three state HMM analysis). Third, to exclude genes with active unannotated TSSs in the analysis region, genes were rejected if the TSS-proximal signal was not more than 10 times the distal signal. Lastly, only genes with Bru-seq expression above 0.5 RPKM were used in the analysis.

6.5.4 Data processing and normalization for elongation rate analysis

The genomic distance analyzed for each transcript extended from 10 kb upstream from the TSS to the minimum acceptable transcript length in that analysis (40 kb in the three state HMM analysis or 150 kb for the four state HHM analysis). This distance was divided into 250 bp bins and the reads along these bins were used to determine the RPKM value of each bin. To minimize the effect of any potential background contamination of unlabeled mature RNA on the elongation rate determinations, the expression signal of bins that overlapped exons was replaced by an interpolation based on the signal of the adjacent bins that did not overlap exons. In order to limit the effect of the transcript’s expression value in the elongation rate analysis, the data was quantile normalized using the R package preprocessCore (Bol-

stad et al., 2003). Since most of the expression signal accumulated in the advancing and receding waves, there were a large number of bins that presented very low expression values and the distribution of binned expression values in the analysis region was similar to a Gamma distribution. In order to improve the presentation of data to downstream analyses, a z-score Gamma- equivalent normalization was carried out using the R package limma (Smyth et al., 2005).

6.5.5 Hidden Markov Model for elongation rate analysis

A Hidden Markov Model was used to determine the elongation rate of each transcript. This analysis was carried out in two different ways. In the first analysis, the position of three states was predicted in genes that were 40 kb or longer. State 1 represented the low signal region upstream from TSS; State 2 represented the advancing wave with high transcription signal; State 3 represented the low signal region downstream from the advancing elongation wave (Fig. 6.1B). In the second analysis, a fourth state was added representing the receding wave (Fig. 6.2D). This second analysis was applied to genes that were at least 150 kb long. Each gene analysis region was split into 250 bp bins and bin RPKM values were calculated from the BruDRB-seq samples. The expression values were normalized (see "Data processing and normalization for elongation rate analysis") and used as the observed layer in the model. The model was trained on regions that were observed to behave as the desired states in the aggregate view of the data (Fig. 6.1B). The relative bin positions used to calculate the output probabilities were: (state 1) from 10 kb to 0.5 kb upstream from the TSS; (state 2) from 0.5 kb to 20 kb downstream from the TSS; (state 3) from 40 kb to 60 kb (four state analysis) or 30 kb to 40 kb downstream (three state analysis) from the TSS; (state 4) from 120 kb to 150 kb downstream from the TSS. The normalized expression values observed in each bins within the

described ranges were pooled and used to determine the emission probabilities for each state. The model was set up so that transitions could only occur from state 1 to state 2, state 2 to state 3, and state 3 to state 4 (when analyzing long genes). The transition probabilities between these states were set to 0.00001.

The emission and transition probabilities were used to fit the multi-state HMM to the data for the complete analysis region for each transcript using the R package msm (Jackson, 2011). The most likely state of each bin was estimated using the Viterbi algorithm. Transcripts where the advancing wave (state 2) began more than 2 kb upstream or downstream from the annotated TSS were removed. Transcripts where the trough (state 3) began at the annotated TSS and transcripts where a state 3 or state 4 (in the long gene analysis) was not recognized were removed from the analysis.

6.5.6 Clustering of genes according to elongation rate

In order to compare between cell lines, the measured elongation rates were quantile normalized. A dissimilarity matrix was then calculated from the quantile normalized elongation rates using an Euclidean distance. This metric was used for the clustering, which was carried out using the k-medoids algorithm (also known as partitioning around medoids, or PAM). The dissimilarity matrix calculation and clustering were performed using the R package cluster (Maechler et al., 2013).

6.5.7 Enrichment of gene sets according to elongation rate

The tool Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005) was used to determine if there were gene sets enriched among the genes with higher or lower elongation rates. The gene set collections used were: positional, curated (BioCarta and KEGG), gene ontology (biological processes, cellular components

and molecular functions), and oncogenic signatures (downloaded from the Molecular Signatures Database (MSigDB) version 4; `http://www.broadinstitute.org/gsea/msigdb/index.jsp`). GSEA was run on a list of genes ranked according to elongation rate and gene sets with at least 15 represented genes were selected for analysis. A false discovery rate (FDR) adjusted p-values threshold of 0.25 was applied to determine enrichment of a gene set.

6.5.8 Correlation between elongation rate and gene features

To determine if the elongation rates were correlated with different physical properties of the genes such as DNA sequence, several different features were tested in a permutation test. Seven features were analyzed: (1) Transcript length (in base pairs) using the Ensembl annotation (Flicek et al., 2013). (2) Distance to nearby expressed genes (in base pairs). Genes with an expression level greater than 0.1 RPKM were considered expressed. Distances were measured to closest upstream and downstream gene in either the sense or antisense orientation, resulting in a total of four different values. (3) Density of exons. The Ensembl's project exon annotation was used (Flicek et al., 2013). All annotated exons were used and exons that overlapped were merged into a single exon. (4) GC content. (5) Repetitive DNA (combined length of each class in base pairs). The RepeatMasker annotation (`http://www.repeatmasker.org/`) was downloaded from the UCSC genome browser (`http://genome.ucsc.edu/`). The repetitive DNA annotation was simplified to reflect only the major classes (i.e. DNA, LINE, low complexity, LTR, other, RC/Helitron, RNA, rRNA, satellite, scRNA, simple repeat, SINE, snRNA, srpRNA, tRNA, unknown). Only non-overlapping repetitive regions were used in the analysis. (6) Non-B DNA (combined length of each class in base pairs). The Non- B DB v2.0 annotation was used in this analysis (Cer et al.,

2013). The classes of non-B DNA used were: A phased repeat, direct repeat, G-quadruplex motif, inverted repeat, mirror repeat, short tandem repeat, and Z DNA motif (7) Density of methylated CpG sites. The DNA Methylation by Reduced Representation Bisulfite-seq from ENCODE/HudsonAlpha dataset was used (`http://genome.ucsc.edu/cgibin/hgFileUi?db=hg19&g=wgEncodeHaibMethylRrbs`). The tracks used (and their respective GEO accession ID) were: K562 HudsonAlpha replicates 1 (GSM683856) and 2 (GSM68378), and MCF-7 Standford replicates 1 (GSM720350) and 2 (GSM720353). A CpG site was only used in the analysis if it was represented in a minimum of 10 reads and if at least 90% of those reads indicated that that site was in fact methylated.

To assess the correlations between high and low levels of elongation and a particular DNA or chromatin feature, we divided the feature metric (e.g. exon count) by the length of elongation (providing a feature density). It was observed that certain features presented a non-random distribution throughout the length of the gene. For example, GC content tends to be higher nearby the TSS and decrease as the distance to the TSS increases until it levels off. Therefore, if one assigned random elongation rates to genes and measured their GC content there would be a negative correlation. Due to this limitation, a permutation test was performed by randomly distributing the observed elongation rates among the genes.

In order to limit the effect of outliers, the 5% most extreme elongation rate values (top and bottom 2.5%) were excluded from the analysis. Also, features were only analyzed if they had been measured in at least 20% of the transcripts. The feature metric was measured for all genes under the new elongation area as determined by the randomly distributed elongation rates. A regression coefficient between elongation rate and feature metric was calculated and stored. This process was repeated 2000

times. Finally, the regression coefficient observed in the original data was compared to the 2000 permutated regression coefficients. A one-tailed p-value was determined by measuring the percentage of times that a permutated regression coefficient was equal to or more extreme than the observed regression coefficient. To account for the multiple testing, the p-values were FDR-corrected. A corrected regression coefficient was calculated by subtracting the observed regression coefficient from the median value of the 2000 permutated regression coefficients.

6.5.9 Long-range promoter interaction and elongation rate

Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) data was used to determine if the elongation rate of genes in contact with the same transcription machinery is correlated. A regression analysis between elongation rates of genes believed to be physically in contact with each other as assessed with ChIA-PET was carried out. The data was downloaded from the UCSC genome browser (`http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeGisChiaPet`). The tracks used (and their respective GEO accession ID) were: K562 POL2 replicates 1 and 2 (GSM970213), and MCF-7 POL2 replicates 3 and 4 (GSM970209). The segments that were considered to be connected by ChIA-PET analysis were intersected with the annotation of genes for which elongation rates were measured. Two genes were considered to be physically connected if two connected segments overlapped the TSS of the two genes.

6.5.10 Aggregate signal of ChIP-seq data for the elongation rate quartiles

To determine if transcription elongation rates were correlated to the density of specific histone modifications or proteins, we downloaded the ChIP-seq processed signal files (in the bigWig file format) from the UCSC genome browser. Data was obtained

from `http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeBroadHistone` and `http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeSydhTfbs` datasets. The tracks used (and their respective GEO accession ID) were: ChIP-seq input control (GSM733780), H3K4me1 (GSM733692), H3k4me3 (GSM733680), H3K27ac (GSM733656), H3K9me1 (GSM733777), H3K9me3 (GSM733776), H3K9ac (GSM733778), H3K27me3 (GSM733658), H3K36me3 (GSM733714), H3K79me2 (GSM733653), H4K20me1 (GSM733675), CTCF (GSM733719), Pol2 (GSM733643), CCNT2 (GSM935547), GTF21 (GSM935501), NELF-E (GSM935392), cMyc (GSM935516). This data was normalized according to (Ram et al., 2011).

A genomic region encompassing 5 kb upstream to 20 kb downstream of the TSS of all genes for which an elongation rate was recorded was used in the analysis. This analysis region was split into bins of 250 bp in length. For each bin, the average ChIP- seq signal of a given data set was calculated and plotted according to each quadrant of elongation rates.

6.5.11 Data access

All the primary sequencing data files used in this study have been deposited in NCBIs Gene Expression Omnibus (GEO) with the accession number GSE55534.

6.6 Acknowledgments

We thank Manhong Dai and Fan Meng for administration and maintenance of the University of Michigan Molecular and Behavioral Neuroscience Institute (MBNI) computing cluster and the personnel at the University of Michigan Sequencing Core for technical assistance. We also thank all members of the Ljungman lab for valuable discussions. This work was supported by funds from National Institute of Environmental Health Sciences (1R21ES020946), National Human Genome Research

Institute (1R01HG006786) and the National Institute of Health (P50CA130810).

Table 6.1: Correlations between DNA or genomic features and transcription elongation rates. (+) Positive correlation, (-) negative correlation, (NS) not statistically significant, (NA) not assessed.

Feature	HF1	TM	CS-B	K562	MCF7
Exon density	-	-	-	-	-
GC content	-	NS	-	-	-
Long terminal repeats	-	NS	-	-	NS
CpG methylation	NA	NA	NA	-	-
Low complexity sequences	+	+	+	+	+
Gene length	+	+	+	+	+
Distance from nearby transcription unit	+	+	+	+	+

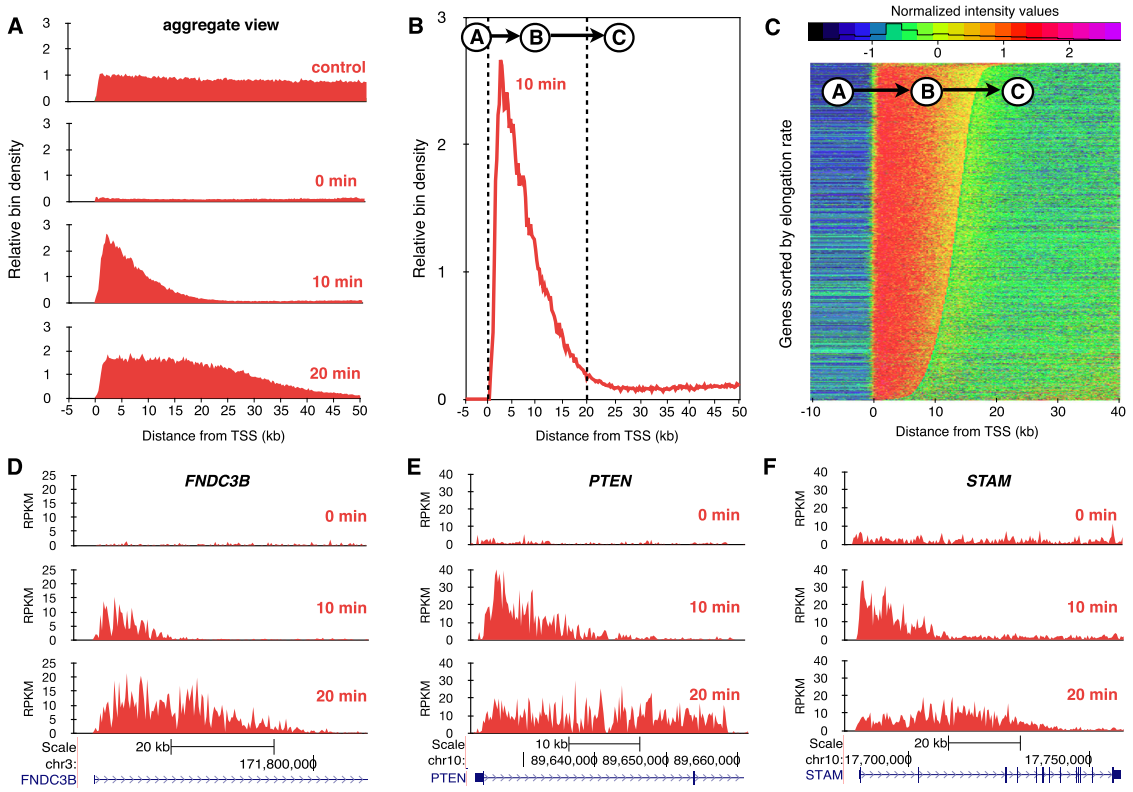


Figure 6.1: Transcription elongation rates measured genome-wide using BruDRB-seq. (A) Aggregate view of nascent RNA reads through the first 50 kb of large expressed genes in the human fibroblast cell line HF1. Control: Bru labeling for 30 min; 0 min: Bru labeling during the last 10 min of a 60 minute DRB treatment; 10 min: appearance of a nascent transcription wave at the 5'-end of genes during a 10 minute recovery after DRB removal (10 min Bru labeling during recovery period); 20 min: advancing nascent transcription wave after a 20 minute recovery time following DRB removal (Bru labeling during last 10 min of recovery). (B) Aggregate view of BruDRB-seq (10' recovery) showing the (A), upstream region of TSS having a low signal; (B), the advancing wave and (C), region downstream of the advancing wave with low signal. (C) A Hidden Markov Model was developed to identify advancing waves and measure their lengths, which are proportional to their elongation rates having A, B, and C represent the three states of this model. Normalized signal of genes in HF1 cells ordered by elongation rate for both a 10- and 20-minute recovery following DRB removal are shown. Examples of transcriptional recovery in individual genes after 0, 10, and 20 min recovery after DRB removal in HF1 cells are shown in (D), (E), and (F).

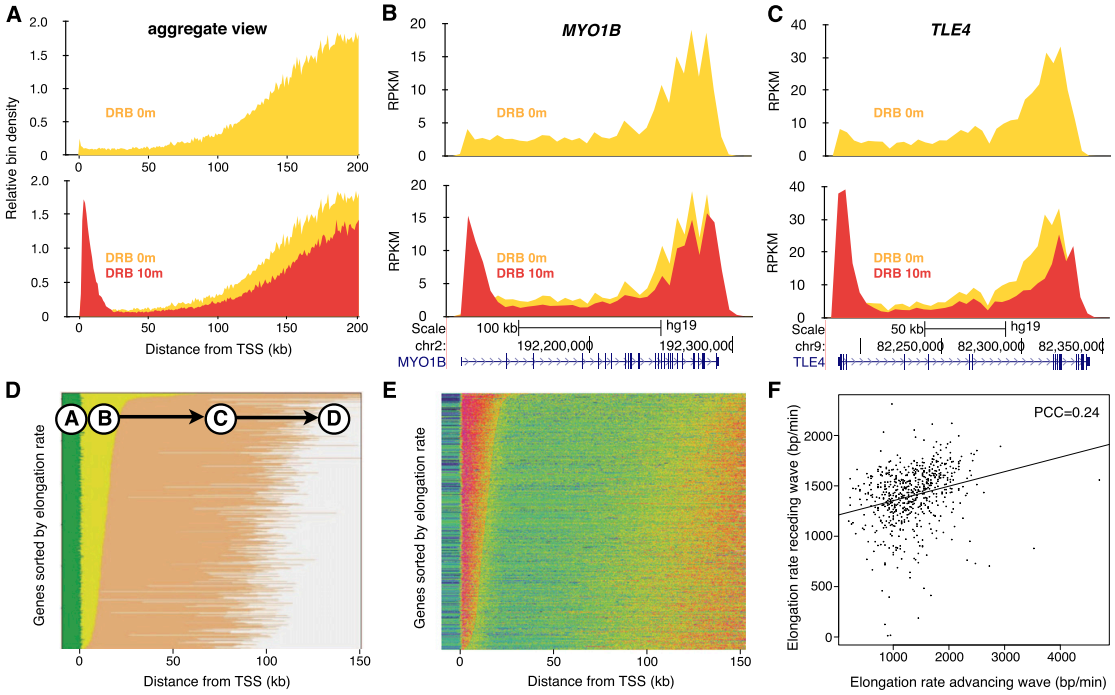


Figure 6.2: The relationship between advancing and receding transcription elongation waves. (A) Aggregate view of 189 genes longer than 200 kb in HF1 cells during DRB treatment (yellow) or following a 10 minute recovery period after DRB treatment (red). (B) Advancing and receding transcription waves in the large genes MYO1B and (C) TLE4. (D) A 4-state Hidden Markov Model was developed to take into account this receding wave in large genes. Genes in HF1 cells were ordered by the length of the advancing wave (state B) and pseudo-colored by state (green: A, yellow: B, orange: C, gray: D). (E) Normalized signaling for genes ordered according to the length of state B. (f) The relationship between elongation rates calculated by the length of the advancing wave (state B) and the distance of the trough between the end of the advancing wave and the beginning of the receding wave (state C). PCC, Pearson's Correlation Coefficent.

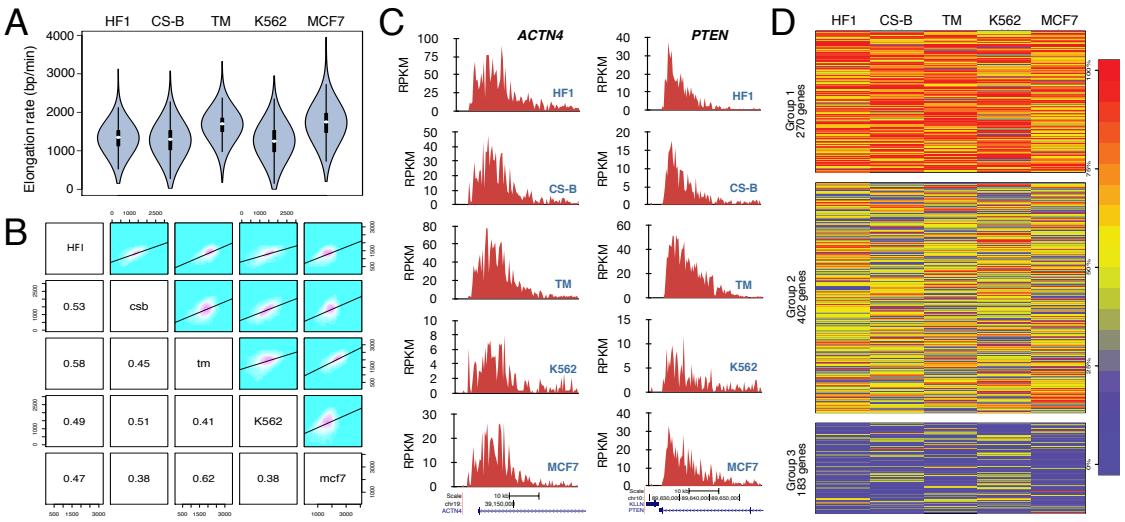


Figure 6.3: Comparisons of transcription elongation rates among five cell lines. (A) Cells were treated with DRB for 60 min followed by drug reversal and immediate incubation with 2 mM Bru for 10 min. BruDRB-seq was then performed and violin plots illustrating the distribution of elongation rates in the indicated cell lines are shown (the interquartile ranges are represented by thick vertical bars and white dots indicate the median values). Sample sizes: HF1 - 2702, CS-B - 1932, TM - 2469, K562 - 2270, MCF-7 - 2399. (B) Grid of pairwise comparisons of elongation rates between each of five cell lines. Each individual comparison includes those genes with measurable elongation rates that are expressed in both cell lines. Frequencies and linear regression models are plotted in the upper-right panels and respective Pearsons correlation coefficients in the lower-left. (C) Examples of two individual genes showing similar elongation rates in the 5 cell lines. (D) Genes expressed in all five cell lines (855 genes) were clustered by normalized elongation rate into 3 groups using the k-medoids method. Genes in Group 1 tend to be faster-elongating in multiple cell lines and genes in Group 3 tend to be slower in multiple cells. Genes belonging to Group 2 (47% of total genes) consists of genes with intermediate or variable elongation rates across cell lines. Genes are colored by percentile ranking within each cell line (100%=highest elongation rate=red). HF1: human foreskin fibroblasts; CS-B: Cockayne syndrome fibroblasts B; TM: human skin fibroblasts; K562: myelogenous leukemia; MCF-7: breast cancer.

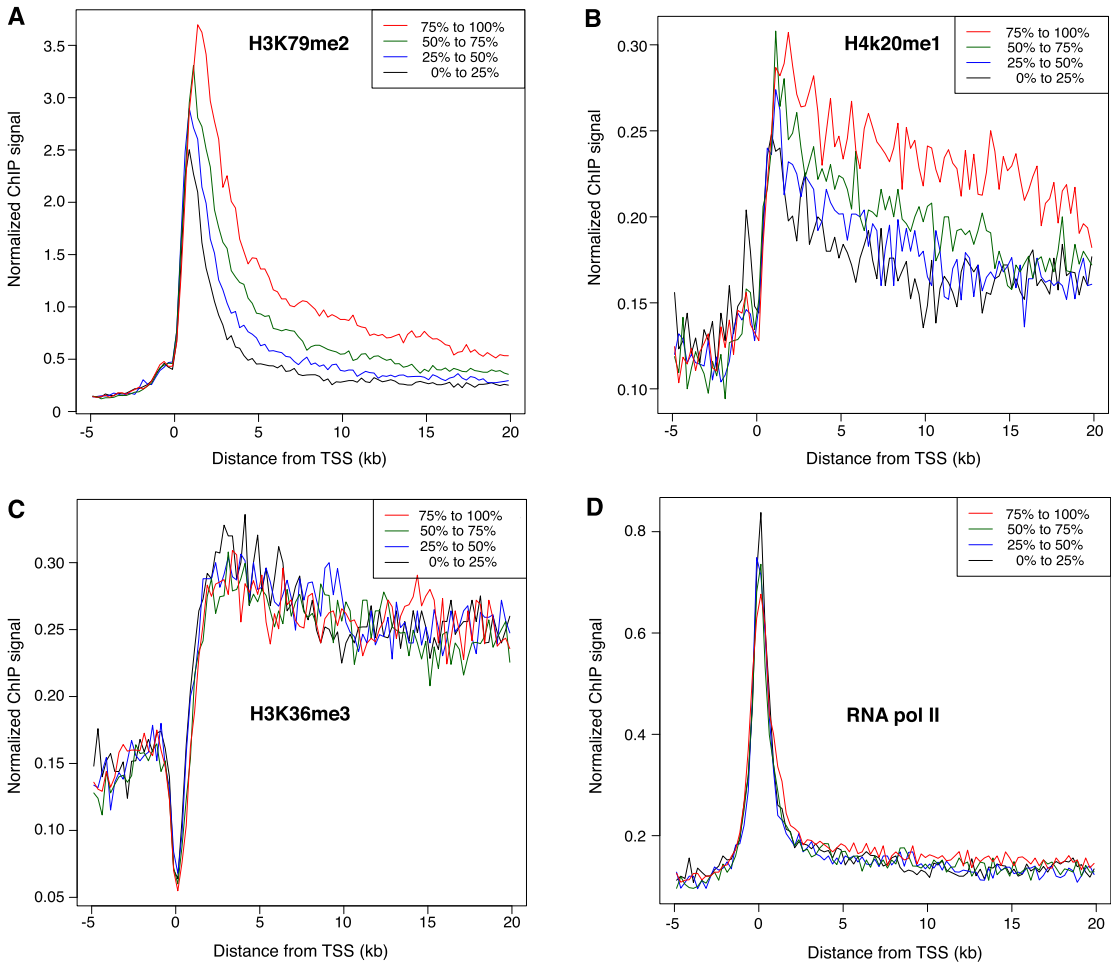


Figure 6.4: Elongation rates are associated with specific histone modifications. Genes expressed in K562 cells were ranked according to elongation rate and placed into four equal-sized groups (from fastest to slowest: red, green, blue, black). ChIP-seq data for K562 cells was obtained from ENCODE and median binned values for each group plotted as indicated for (A) H3K79me2, (B) H4K20me1, (C) H3K36me3, and (D) RNA polymerase II. In (A) and (B), genes with faster elongation rates have a higher density of histone modification both near the transcription start site (TSS) and within the gene bodies. In (C) and (D), neither histone modification nor RNA polymerase II occupancy correlated to transcription elongating rates.

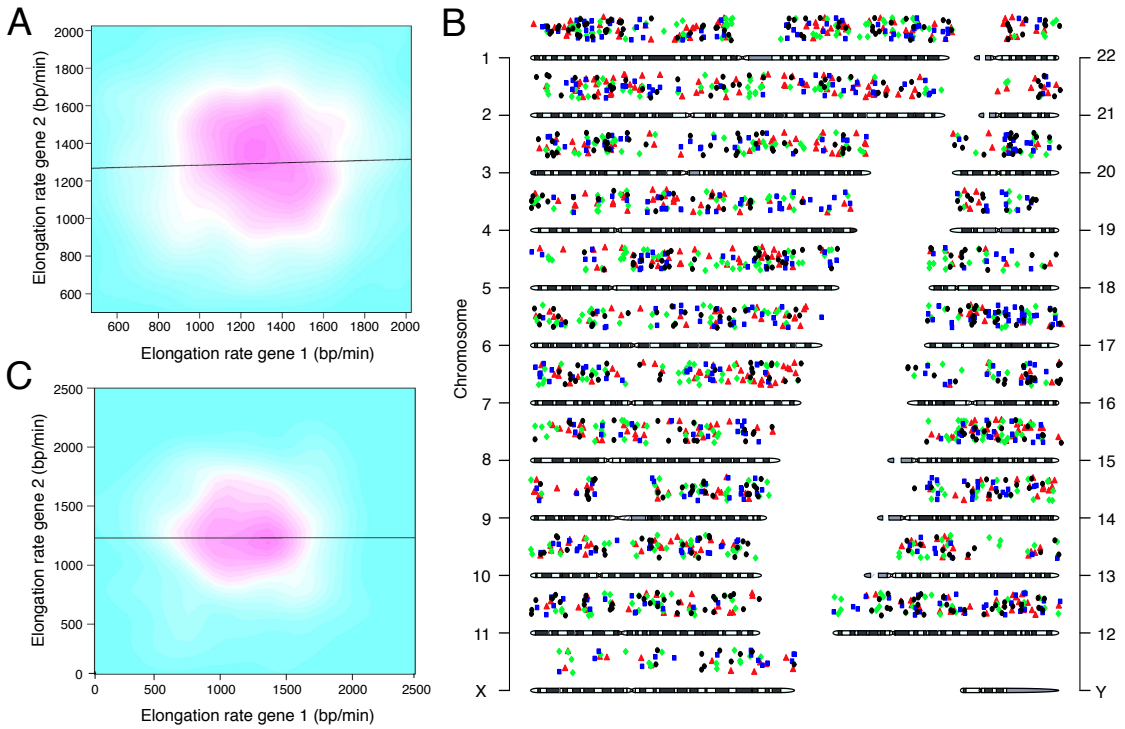


Figure S6.1: Transcription elongation rates in K562 cells are not related to two- or three-dimensional localization. (A) Correlation between the elongation rate of a particular gene and the elongation rate of its nearest expressing neighbor. Data point frequency illustrated by a colored contour plot where solid line indicates a linear regression model. (B) Elongation rates of genes in relationship their two-dimensional chromosomal location. Genes were divided into four equal-sized groups according to ranked elongation rates (from slowest to fastest: black circles, blue squares, green diamonds, red triangles) and their chromosomal location are denoted for each chromosome. (C) Correlation between elongation rates and three-dimensional association using ChIA-PET data for K562 cells from ENCODE. Data point frequency illustrated by a colored contour plot where solid line indicates a linear regression model.

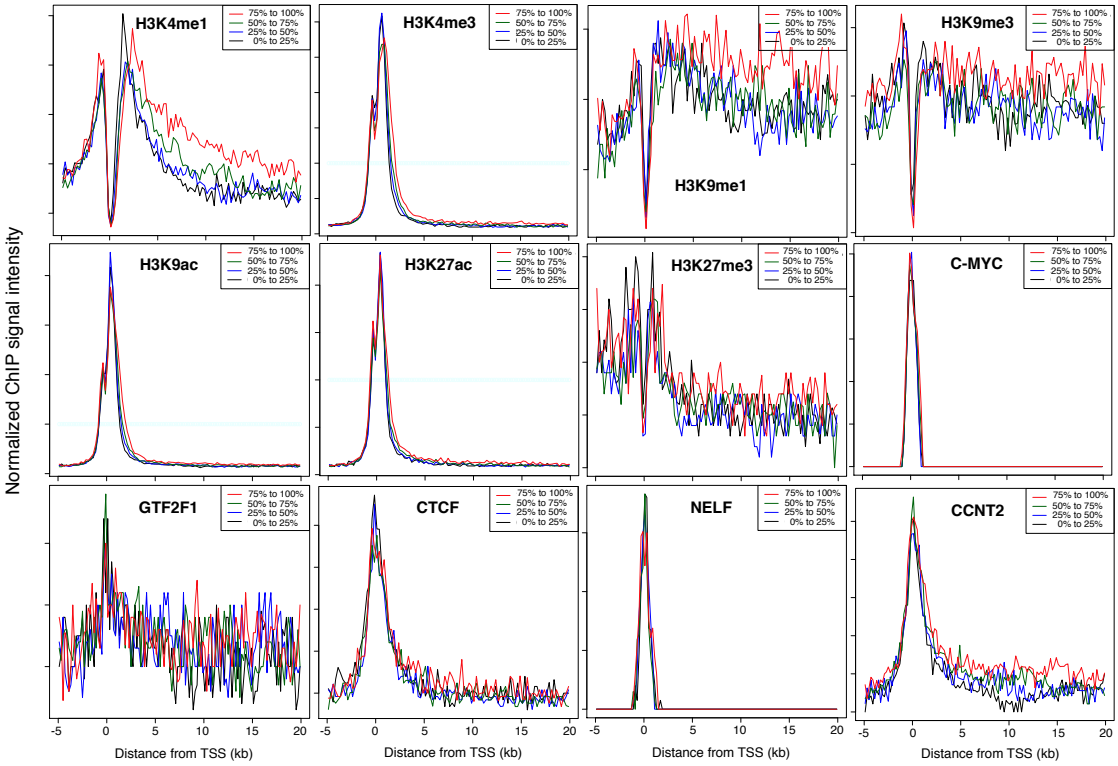


Figure S6.2: Comparisons between elongation rate and histone modification or transcription factor binding. Genes expressed in K562 cells were ranked according to elongation rate and placed into four equal-sized groups (from fastest to slowest: red, green, blue, black). CHiP-Seq signal was obtained from ENCODE and median binned values for each group were plotted as indicated.

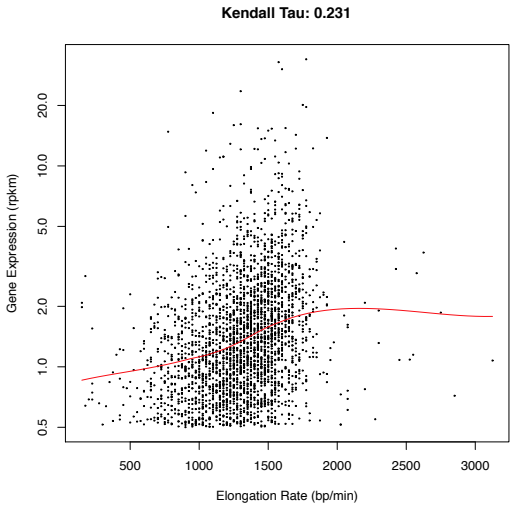


Figure S6.3: Relationship between transcription elongation rate and nascent RNA transcription (measured using Bru-seq in the 40kb analysis region) for the HF1 cell line. The red trace indicates a local regression (LOESS) fit and the title indicates Kendall's Tau correlation coefficient.

CHAPTER VII

Concluding remarks

RNA transcription, processing and degradation are extremely complex and intertwined processes. While most modern transcriptomic studies have been built upon the final product of these processes, there is still a clear necessity to better understand the steps that lead up to mature RNA. In this thesis I have described a suite of techniques designed to investigate different stages of RNA transcription. These techniques are based on the metabolic labeling of nascent RNA with bromouridine (thoroughly described on Chapters II and III). This allows us to specifically capture and measure the RNA produced during the labeling period.

In response to environmental challenges or stimuli, cells undergo a re-programming of gene expression. In Chapter III we use Bru-seq and BruChase-seq to interrogate the effects the pro-inflammatory TNF protein have on both synthesis and stability of RNA genome-wide. The cells were exposed to TNF for one hour prior to nascent RNA labeling, and labeled for the next 30 minutes in the presence of TNF. As it would be expected, we noticed gene synthesis changes that were consistent with an inflammatory response (section 3.3.7). One of the most interesting aspects of a visual inspection of the gene expression changes was identifying genes where the post-induction transcription level was not constant throughout the gene (see figure

S3.1).

Changes in gene synthesis are caused by a change in the number of polymerases initiating transcription for that given gene. The change in nascent RNA, therefore, should initially be visible directly downstream to the promoter and, with time, "travel" through the rest of the gene. Given enough time and a constant initiation rate, the nascent transcription signal of a gene should reach a reasonably constant level throughout the gene. The fact that we observed genes where this constant level had not been reached after 1 hour of TNF exposure played an important role in changing how I think about the genome.

Clearly, one of the main reasons why the signal in the genes in figure S3.1 has not yet plateaued is their size. These genes are close to 200kb long, which is larger than the genome size of certain bacteria (Bennett and Moran, 2013). The presence of such large genes in the human genome certainly puts nascent transcription in perspective. Treatment induced changes in gene synthesis can only affect the total level of RNA once that gene is completely transcribed. Therefore, transcription elongation is an extremely important process regulating RNA production.

In order to further explore RNA synthesis we decided to block transcription elongation. This was carried out in two different ways. In chapter IV, we used UV radiation to induce transcription elongation inhibiting lesions to the DNA (see section 1.5.2). In chapter V, the anti-cancer drug camptothecin was used to inhibit the functioning of topoisomerase I, which leads to the creation of barriers for elongating RNAPII (see section 1.5.3). Since both these treatments blocked transcription elongation, their results displayed some very interesting similarities. For example, after treatment increased signal was observed in shorter genes (figures 5.1C and 4.1). Also, signal was redistributed into promoters (figures 5.3 and 4.2) and putative enhancer

elements figures 5.2C and 4.5. This indicates that inhibition of transcription elongation leads to similar signal redistribution regardless of the inhibition mechanism.

The fact that the effect of transcription elongation inhibitors on gene expression is affected by gene size is not surprising, but it is interesting. After the removal of the transcription elongation blocking agent, our data suggested that transcription needs to restart from the promoter (figures 5.3 and 5.4). This has a particularly strong effect on large genes, since they take longer to be fully transcribed. Consider a gene whose transcription takes 2 hours. If it is blocked at the elongation stage after 1 hour and 50 minutes, the incomplete transcript will probably be degraded and transcription will have to re-initiate. This could be extremely problematic for the cell, depending on the importance of the gene for the cell's overall health.

A major aspect in determining how long it takes for a gene to be transcribed is the elongation rate of RNAPII. In chapter VI, we used nascent RNA transcription to calculate genome-wide elongation rates. We found that the abundance of several sequence patterns was correlated to the elongation rate of RNAPII. For example, the density of long terminal repeats was negatively correlated with elongation rates (see table 6.1). Therefore, the sequence composition traversed by RNAPII seems to affect its elongation rate. More interestingly, we noticed that the abundance of certain epigenetic modifications was correlated to elongation rate (see figure 6.4). If a causal relationship exists between these epigenetic modifications and elongation rate, it would be reasonable to assume that a cell is capable of modifying a gene's elongation rate. This could be used as a mechanism to regulate gene expression timing. The most surprising finding, however, was that the gene's length and distance to other gene's was positively correlated to elongation rate.

It is hard to imagine how these two factors could affect elongation rate. Assum-

ing non-overlapping genes, a very large gene is, by definition, far away from other genes transcribing downstream from it. Therefore, it is possible that the main factor determining this correlation is the distance to other transcribing genes. The advantage of being isolated from other transcribing genes could possibly result from an inter-gene competition for resources such as nucleotides. Regardless, the differences in elongation rate found within transcripts is not extremely large 6.3. Therefore, the transcription of a very large gene ($> 300kb$) would take much longer than the transcription of a small gene ($< 5kb$) even if their elongation rates were at the extreme of the ranges observed in Chapter VI.

Therefore, all of these findings bring us back to the long genes whose TNF-induced synthesis changes were still unnoticed 1 hour after treatment (figure S3.1). Is it important for the cell that these gene's full transcript is completed long after exposure? Is gene size a mechanism for adjusting the timing of gene expression? There is research that suggests that this could be the case for genes involved in development (Swinburne and Silver, 2008). But how prevalent is this mechanism? Has gene length been optimized by evolution for all transcriptomic responses? This also helps putting gene transcription studies in perspective. When studying treatment induced changes in gene expression at the total RNA level, the size of the gene is crucial to determine when changes in expression can be observed. This is particularly true if the time span after treatment being studied is short.

Time is not only important for the termination of transcription. Treatment induced changes in transcription initiation do no occur at a constant rate through time. As time passes, the transcription initiation rates can change. Figure S3.1C shows an interesting example of a gene that was upregulated in the initial stages of the response to TNF. Later, however, its expression returns to pre-TNF levels. In most

studies, the expression value for that gene would be calculated by integrating all signal observed throughout the gene. This approach would basically average out the gene's transcription initiation rate for the whole treatment period. Clearly, this is a simplistic approach that overlooks the complexity of the gene expression program initiated by TNF exposure.

Therefore, I postulate that a study of gene expression changes is only exhaustive if one investigates changes both in transcription initiation and in mature RNA production at a given time point. This could be accomplished by modifying nascent RNA sequencing techniques, such as Bru-seq, to measure these different steps of transcription. For example, the amount of transcription initiation could be calculated based on the BruUV-seq signal. As shown in figure 4.4, BruUV-seq can be used to measure treatment induced changes in gene expression at gene's promoters. Measuring mature RNA production could be accomplished by initiating metabolic labeling at the same time as a treatment and analyzing the RNA that has been labeled and polyadenylated. In a similar approach, Rabani et al. (2011) studied changes in gene expression in mouse dendritic cells exposed to lipopolysaccharide in nascent and total RNA. Since they used RNA-seq, however, their measurement of mature RNA could be highly influenced by changes in transcript stability and did not directly reflect how much RNA was produced after the beginning of the treatment.

Overall, the research carried out in this thesis have not only clarified important biological mechanisms, but also demonstrated the complexity of RNA transcription regulation through synthesis and stability. Hopefully these techniques will support further scientific advances and lead to a better understanding of RNA transcription as whole.

BIBLIOGRAPHY

Adelman, K. and Lis, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet*, 13(10):720–731, 2012. doi:10.1038/nrg3293. 9

Aggarwal, B. B. Signalling pathways of the TNF superfamily: a double-edged sword. *Nat Rev Immunol*, 3(9):745–756, 2003. doi:10.1038/nri1184. 73

Aggarwal, B. B., Gupta, S. C. and Kim, J. H. Historical perspectives on tumor necrosis factor and its superfamily: 25 years later, a golden journey. *Blood*, 119(3):651–665, 2012. doi:10.1182/blood-2011-04-325225. 16

Ahn, S. H., Kim, M. and Buratowski, S. Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3' end processing. *Mol Cell*, 13(1):67–76, 2004. 11

Anders, S. and Huber, W. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106, 2010. doi:10.1186/gb-2010-11-10-r106. 46, 71

Anderson, P. Post-transcriptional regulons coordinate the initiation and resolution of inflammation. *Nat Rev Immunol*, 10(1):24–35, 2010. doi:10.1038/nri2685. 17, 63, 70, 73

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F. O., Jrgensen, M., Andersen, P. R., Bertin, N., Rackham, O., Burroughs, A. M., Baillie, J. K., Ishizu, Y., Shimizu, Y., Furuhata, E., Maeda, S., Negishi, Y., Mungall, C. J., Meehan, T. F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C. O., Heutink, P., Hume, D. A., Jensen, T. H., Suzuki, H., Hayashizaki, Y., Müller, F., , F. A. N. T. O. M. C., Forrest, A. R. R., Carninci, P., Rehli, M. and Sandelin, A. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, 2014. doi:10.1038/nature12787. 92, 100, 103

Andrews, S. *FastQC: A quality control tool for high throughput sequence data*, 2010. 21

Anglin, J. L. and Song, Y. A medicinal chemistry perspective for targeting histone H3 lysine-79 methyltransferase DOT1L. *J Med Chem*, 56(22):8972–8983, 2013. doi:10.1021/jm4007752. 145

Ardehali, M. B., Yao, J., Adelman, K., Fuda, N. J., Petesch, S. J., Webb, W. W. and Lis, J. T. Spt6 enhances the elongation rate of rna polymerase II in vivo. *EMBO J*, 28(8):1067–1077, 2009. doi:10.1038/emboj.2009.56. 137

Asin-Cayuela, J. and Gustafsson, C. M. Mitochondrial transcription and its regulation in mammalian cells. *Trends Biochem Sci*, 32(3):111–117, 2007. doi:10.1016/j.tibs.2007.01.003. 69

Baumann, M., Pontiller, J. and Ernst, W. Structure and basal transcription complex

of RNA polymerase II core promoters in the mammalian genome: an overview. *Mol Biotechnol*, 45(3):241–247, 2010. doi:10.1007/s12033-010-9265-6. 7

Beck, D. B., Oda, H., Shen, S. S. and Reinberg, D. PR-Set7 and H4K20me1: at the crossroads of genome integrity, cell cycle, chromosome condensation, and transcription. *Genes Dev*, 26(4):325–337, 2012. doi:10.1101/gad.177444.111. 145

Beg, A. A. and Baltimore, D. An essential role for NF-kappaB in preventing TNF-alpha-induced cell death. *Science*, 274(5288):782–784, 1996. 73

Belotserkovskii, B. P., Mirkin, S. M. and Hanawalt, P. C. DNA sequences that interfere with transcription: implications for genome function and stability. *Chem Rev*, 113(11):8620–8637, 2013. doi:10.1021/cr400078y. 139, 143

Bennett, G. M. and Moran, N. A. Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a Phloem-feeding insect. *Genome Biol Evol*, 5(9):1675–1688, 2013. doi:10.1093/gbe/evt118. 164

Blumenthal, T. Operons in eukaryotes. *Brief Funct Genomic Proteomic*, 3(3):199–211, 2004. 96

Bolstad, B. M., Irizarry, R. A., Astrand, M. and Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003. 147

Bowman, K. J., Newell, D. R., Calvert, A. H. and Curtin, N. J. Differential effects of the poly (ADP-ribose) polymerase (PARP) inhibitor NU1025 on topoisomerase i and ii inhibitor cytotoxicity in L1210 cells in vitro. *Br J Cancer*, 84(1):106–112, 2001. doi:10.1054/bjoc.2000.1555. 125

Brogna, S. and Wen, J. Nonsense-mediated mRNA decay (NMD) mechanisms. *Nat Struct Mol Biol*, 16(2):107–113, 2009. doi:10.1038/nsmb.1550. 51

Cai, H. and Luse, D. S. Transcription initiation by RNA polymerase II in vitro. properties of preinitiation, initiation, and elongation complexes. *J Biol Chem*, 262(1):298–304, 1987. 13

Carswell, E. A., Old, L. J., Kassel, R. L., Green, S., Fiore, N. and Williamson, B. An endotoxin-induced serum factor that causes necrosis of tumors. *Proc Natl Acad Sci U S A*, 72(9):3666–3670, 1975. 16

Cer, R. Z., Donohue, D. E., Mudunuri, U. S., Temiz, N. A., Loss, M. A., Starner, N. J., Halusa, G. N., Volfovsky, N., Yi, M., Luke, B. T., Bacolla, A., Collins, J. R. and Stephens, R. M. Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res*, 41(Database issue):D94–D100, 2013. doi:10.1093/nar/gks955. 150

Chapman, R. D., Heidemann, M., Albert, T. K., Mailhammer, R., Flatley, A., Meisterernst, M., Kremmer, E. and Eick, D. Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7. *Science*, 318(5857):1780–1782, 2007. doi:10.1126/science.1145977. 9

Chen, A. Y., Choy, H. and Rothenberg, M. L. DNA topoisomerase I-targeting drugs as radiation sensitizers. *Oncology (Williston Park)*, 13(10 Suppl 5):39–46, 1999. 125

Churchman, L. S. and Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, 469(7330):368–373, 2011. doi:10.1038/nature09652. 11, 14, 29, 63

Clark, M. B., Johnston, R. L., Inostroza-Ponta, M., Fox, A. H., Fortini, E., Moscato, P., Dinger, M. E. and Mattick, J. S. Genome-wide analysis of long noncoding RNA stability. *Genome Res*, 22(5):885–898, 2012. doi:10.1101/gr.131037.111. 68

Close, P., East, P., Dirac-Svejstrup, A. B., Hartmann, H., Heron, M., Maslen, S., Chariot, A., Söding, J., Skehel, M. and Svejstrup, J. Q. DBIRD complex integrates alternative mrna splicing with rna polymerase ii transcript elongation. *Nature*, 484(7394):386–389, 2012. doi:10.1038/nature10925. 133

Core, L. J., Waterfall, J. J., Gilchrist, D. A., Fargo, D. C., Kwak, H., Adelman, K. and Lis, J. T. Defining the status of RNA polymerase at promoters. *Cell Rep*, 2(4):1025–1035, 2012. doi:10.1016/j.celrep.2012.08.034. 13, 28, 91

Core, L. J., Waterfall, J. J. and Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909):1845–1848, 2008. doi:10.1126/science.1162228. 9, 12, 28, 29, 63, 64, 91

Dani, C., Blanchard, J. M., Piechaczyk, M., El Sabouty, S., Marty, L. and Jeanteur, P. Extreme instability of myc mRNA in normal and transformed human cells. *Proc Natl Acad Sci U S A*, 81(22):7046–7050, 1984. 50

Danko, C. G., Hah, N., Luo, X., Martins, A. L., Core, L., Lis, J. T., Siepel, A. and Kraus, W. L. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol Cell*, 50(2):212–222, 2013. doi:10.1016/j.molcel.2013.02.015. 13, 24, 28, 121, 134, 135, 137, 142

Darzacq, X., Shav-Tal, Y., de Turris, V., Brody, Y., Shenoy, S. M., Phair, R. D. and Singer, R. H. In vivo dynamics of RNA polymerase II transcription. *Nat Struct Mol Biol*, 14(9):796–806, 2007a. doi:10.1038/nsmb1280. 11, 24

Darzacq, X., Shav-Tal, Y., de Turris, V., Brody, Y., Shenoy, S. M., Phair, R. D. and Singer, R. H. In vivo dynamics of RNA polymerase II transcription. *Nat Struct Mol Biol*, 14(9):796–806, 2007b. doi:10.1038/nsmb1280. 134, 142

Day, N., Hemmaplardh, A., Thurman, R. E., Stamatoyannopoulos, J. A. and Noble, W. S. Unsupervised segmentation of continuous genomic data. *Bioinformatics*, 23(11):1424–1426, 2007. doi:10.1093/bioinformatics/btm096. 83, 135

de la Mata, M., Alonso, C. R., Kadener, S., Fededa, J. P., Blaustein, M., Pelisch, F., Cramer, P., Bentley, D. and Kornblihtt, A. R. A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell*, 12(2):525–532, 2003. 11

Dekker, J., Marti-Renom, M. A. and Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet*, 14(6):390–403, 2013. doi:10.1038/nrg3454. 140

Deng, W. and Roberts, S. G. E. TFIIB and the regulation of transcription by RNA polymerase II. *Chromosoma*, 116(5):417–429, 2007. doi:10.1007/s00412-007-0113-9. 7

Desai, S. D., Liu, L. F., Vazquez-Abad, D. and D'Arpa, P. Ubiquitin-dependent destruction of topoisomerase I is stimulated by the antitumor drug camptothecin. *J Biol Chem*, 272(39):24159–24164, 1997. 115

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Rder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T.,

Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guig, R. and Gingeras, T. R. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, 2012. doi:10.1038/nature11233. 68

Dölken, L., Ruzsics, Z., Rdle, B., Friedel, C. C., Zimmer, R., Mages, J., Hoffmann, R., Dickinson, P., Forster, T., Ghazal, P. and Koszinowski, U. H. High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA*, 14(9):1959–1972, 2008. doi:10.1261/rna.1136108. 73

Donahue, B. A., Yin, S., Taylor, J. S., Reines, D. and Hanawalt, P. C. Transcript cleavage by RNA polymerase II arrested by a cyclobutane pyrimidine dimer in the DNA template. *Proc Natl Acad Sci U S A*, 91(18):8502–8506, 1994. 93

Dubois, M. F., Nguyen, V. T., Bellier, S. and Bensaude, O. Inhibitors of transcription such as 5,6-dichloro-1-beta-D-ribofuranosylbenzimidazole and isoquinoline sulfonamide derivatives (h-8 and h-7) promote dephosphorylation of the carboxyl-terminal domain of RNA polymerase II largest subunit. *J Biol Chem*, 269(18):13331–13336, 1994. 136

Durand-Dubief, M., Svensson, J. P., Persson, J. and Ekwall, K. Topoisomerases,

chromatin and transcription termination. *Transcription*, 2(2):66–70, 2011. doi: 10.4161/trns.2.2.14411. 119

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A. and Huber, W. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, 2005. doi:10.1093/bioinformatics/bti525. 146

Egloff, S. and Murphy, S. Cracking the RNA polymerase II CTD code. *Trends Genet*, 24(6):280–288, 2008. doi:10.1016/j.tig.2008.03.008. 9, 10

ENCODE Project Consortium, Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C. and Snyder, M. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012. doi:10.1038/nature11247. 101, 105, 141

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M. and Bernstein, B. E. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011. doi:10.1038/nature09906. 141

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Garca-Girn, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Khri, A. K., Keenan, S., Komorowska, M., Kulesha, E., Longden, I., Maurel, T., McLaren, W. M., Muffato, M., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H. S., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T. J. P., Johnson,

N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A. and Searle, S. M. J. Ensembl 2013. *Nucleic Acids Res*, 41(Database issue):D48–D55, 2013. doi: 10.1093/nar/gks1236. 23, 104, 146, 150

Forbes, S., Clements, J., Dawson, E., Bamford, S., Webb, T., Dogan, A., Flanagan, A., Teague, J., Wooster, R., Futreal, P. A. and Stratton, M. R. Cosmic 2005. *Br J Cancer*, 94(2):318–322, 2006. doi:10.1038/sj.bjc.6602928. 51

Friedberg, E. C., Walker, G. C., Siede, W., Wood, R. D., Schultz, R. A. and Ellenberger, T. *DNA Repair and Mutagenesis*. ASM Press, Washington, D.C, 2006. ISBN 978-1555813192. 17, 93

Fu, H., Maunakea, A. K., Martin, M. M., Huang, L., Zhang, Y., Ryan, M., Kim, R., Lin, C. M., Zhao, K. and Aladjem, M. I. Methylation of histone H3 on lysine 79 associates with a group of replication origins and helps limit DNA replication once per cell cycle. *PLoS Genet*, 9(6):e1003542, 2013. doi:10.1371/journal.pgen.1003542. 145

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., Chew, E. G. Y., Huang, P. Y. H., Welboren, W.-J., Han, Y., Ooi, H. S., Ariyaratne, P. N., Vega, V. B., Luo, Y., Tan, P. Y., Choy, P. Y., Wansa, K. D. S. A., Zhao, B., Lim, K. S., Leow, S. C., Yow, J. S., Joseph, R., Li, H., Desai, K. V., Thomsen, J. S., Lee, Y. K., Karuturi, R. K. M., Herve, T., Bourque, G., Stunnenberg, H. G., Ruan, X., Cacheux-Rataboul, V., Sung, W.-K., Liu, E. T., Wei, C.-L., Cheung, E. and Ruan, Y. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462(7269):58–64, 2009. doi:10.1038/nature08497. 22

Gariglio, P. and Mousset, S. Isolation and partial characterization of a nuclear RNA polymerase - SV40 DNA complex. *FEBS Lett*, 56(1):149–155, 1975. 13

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., Min, R., Alves, P., Abyzov, A., Addleman, N., Bhardwaj, N., Boyle, A. P., Cayting, P., Charos, A., Chen, D. Z., Cheng, Y., Clarke, D., Eastman, C., Euskirchen, G., Frietze, S., Fu, Y., Gertz, J., Grubert, F., Harmanci, A., Jain, P., Kasowski, M., Lacroute, P., Leng, J., Lian, J., Monahan, H., O'Geen, H., Ouyang, Z., Partridge, E. C., Patacsil, D., Pauli, F., Raha, D., Ramirez, L., Reddy, T. E., Reed, B., Shi, M., Slifer, T., Wang, J., Wu, L., Yang, X., Yip, K. Y., Zilberman-Schapira, G., Batzoglou, S., Sidow, A., Farnham, P. J., Myers, R. M., Weissman, S. M. and Snyder, M. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489(7414):91–100, 2012. doi:10.1038/nature11245. 63

Gilmour, D. S. and Lis, J. T. RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells. *Mol Cell Biol*, 6(11):3984–3989, 1986. 9

Glover-Cutter, K., Larochelle, S., Erickson, B., Zhang, C., Shokat, K., Fisher, R. P. and Bentley, D. L. TFIIH-associated Cdk7 kinase functions in phosphorylation of C-terminal domain ser7 residues, promoter-proximal pausing, and termination by RNA polymerase II. *Mol Cell Biol*, 29(20):5455–5464, 2009. doi:10.1128/MCB.00637-09. 10

Grigull, J., Mnaimneh, S., Pootoolal, J., Robinson, M. D. and Hughes, T. R. Genome-wide analysis of mRNA stability using transcription inhibitors and mi-

croarrays reveals posttranscriptional control of ribosome biogenesis factors. *Mol Cell Biol*, 24(12):5534–5547, 2004. doi:10.1128/MCB.24.12.5534-5547.2004. 68

Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. and Young, R. A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130(1):77–88, 2007. doi:10.1016/j.cell.2007.05.042. 144

Haider, S. R., Juan, G., Traganos, F. and Darzynkiewicz, Z. Immunoseparation and immunodetection of nucleic acids labeled with halogenated nucleotides. *Exp Cell Res*, 234(2):498–506, 1997. doi:10.1006/excr.1997.3644. 29, 64

Hao, S. and Baltimore, D. The stability of mRNA influences the temporal order of the induction of genes encoding inflammatory molecules. *Nat Immunol*, 10(3):281–288, 2009. doi:10.1038/ni.1699. 17, 63, 70, 71

Hashimoto, S.-i., Suzuki, Y., Kasai, Y., Morohoshi, K., Yamada, T., Sese, J., Morishita, S., Sugano, S. and Matsushima, K. 5'-end SAGE for the analysis of transcriptional start sites. *Nat Biotechnol*, 22(9):1146–1149, 2004. doi:10.1038/nbt998. 92

Hehlgans, T. and Pfeffer, K. The intriguing biology of the tumour necrosis factor/tumour necrosis factor receptor superfamily: players, rules and the games. *Immunology*, 115(1):1–20, 2005. doi:10.1111/j.1365-2567.2005.02143.x. 17

Hintermair, C., Heidemann, M., Koch, F., Descostes, N., Gut, M., Gut, I., Fenouil, R., Ferrier, P., Flatley, A., Kremmer, E., Chapman, R. D., Andrau, J.-C. and Eick, D. Threonine-4 of mammalian RNA polymerase II CTD is targeted by Polo-like kinase 3 and required for transcriptional elongation. *EMBO J*, 31(12):2784–2797, 2012. doi:10.1038/emboj.2012.123. 10

Hirayoshi, K. and Lis, J. T. Nuclear run-on assays: assessing transcription by measuring density of engaged RNA polymerases. *Methods Enzymol*, 304:351–362, 1999. 13

Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., Giardine, B., Ellenbogen, P. M., Bilmes, J. A., Birney, E., Hardison, R. C., Dunham, I., Kellis, M. and Noble, W. S. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res*, 41(2):827–841, 2013. doi:10.1093/nar/gks1284. 92, 101, 106

Holstege, F. C., Fiedler, U. and Timmers, H. T. Three transitions in the RNA polymerase II transcription complex during initiation. *EMBO J*, 16(24):7468–7480, 1997. doi:10.1093/emboj/16.24.7468. 8

Horibata, K., Saijo, M., Bay, M. N., Lan, L., Kuraoka, I., Brooks, P. J., Honma, M., Nohmi, T., Yasui, A. and Tanaka, K. Mutant Cockayne syndrome group B protein inhibits repair of DNA topoisomerase I-DNA covalent complex. *Genes Cells*, 16(1):101–114, 2011. doi:10.1111/j.1365-2443.2010.01467.x. 116, 122

Houseley, J. and Tollervey, D. The many pathways of RNA degradation. *Cell*, 136(4):763–776, 2009. doi:10.1016/j.cell.2009.01.019. 28

Hsiang, Y. H., Lihou, M. G. and Liu, L. F. Arrest of replication forks by drug-stabilized topoisomerase I-DNA cleavable complexes as a mechanism of cell killing by camptothecin. *Cancer Res*, 49(18):5077–5082, 1989. 115

Hsiang, Y. H. and Liu, L. F. Identification of mammalian DNA topoisomerase i as an intracellular target of the anticancer drug camptothecin. *Cancer Res*, 48(7):1722–1726, 1988. 18, 115

Jackson, C. Multi-State Models for Panel Data: The msm Package for R. *Journal of Statistical Software*, 38(8):1–28, 2011. ISSN 1548-7660. 25, 105, 149

Jiang, S., Zhang, L.-F., Zhang, H.-W., Hu, S., Lu, M.-H., Liang, S., Li, B., Li, Y., Li, D., Wang, E.-D. and Liu, M.-F. A novel miR-155/miR-143 cascade controls glycolysis by regulating hexokinase 2 in breast cancer cells. *EMBO J*, 31(8):1985–1998, 2012. doi:10.1038/emboj.2012.45. 72

Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*, 13(7):484–492, 2012. doi:10.1038/nrg3230. 141

Jonkers, I., Kwak, H. and Lis, J. T. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife*, 3:e02407, 2014. 13

Juven-Gershon, T., Hsu, J.-Y., Theisen, J. W. and Kadonaga, J. T. The RNA polymerase II core promoter - the gateway to transcription. *Curr Opin Cell Biol*, 20(3):253–259, 2008. doi:10.1016/j.ceb.2008.03.003. 7

Khabar, K. S. A. Post-transcriptional control during chronic inflammation and cancer: a focus on AU-rich elements. *Cell Mol Life Sci*, 67(17):2937–2955, 2010. doi:10.1007/s00018-010-0383-x. 63, 70

Khodor, Y. L., Rodriguez, J., Abruzzi, K. C., Tang, C.-H. A., Marr, M. T. and Rosbash, M. Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes Dev*, 25(23):2502–2512, 2011. doi:10.1101/gad.178962.111. 14, 29, 63, 65, 80

Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu,

Y., Green, R. D. and Ren, B. A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876–880, 2005. doi:10.1038/nature03877. 9

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G. and Greenberg, M. E. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187, 2010. doi:10.1038/nature09033. 92, 100, 101

Kim, U., Wang, Y., Sanford, T., Zeng, Y. and Nishikura, K. Molecular cloning of cDNA for double-stranded RNA adenosine deaminase, a candidate enzyme for nuclear RNA editing. *Proc Natl Acad Sci U S A*, 91(24):11457–11461, 1994. 14

Kireeva, M. L., Komissarova, N., Waugh, D. S. and Kashlev, M. The 8-nucleotide-long RNA:DNA hybrid is a primary stability determinant of the RNA polymerase II elongation complex. *J Biol Chem*, 275(9):6530–6536, 2000. 11

Komarnitsky, P., Cho, E. J. and Buratowski, S. Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes Dev*, 14(19):2452–2460, 2000. 8

Komissarova, N., Becker, J., Solter, S., Kireeva, M. and Kashlev, M. Shortening of RNA:DNA hybrid in the elongation complex of RNA polymerase is a prerequisite for transcription termination. *Mol Cell*, 10(5):1151–1162, 2002. 11

Kuehner, J. N., Pearson, E. L. and Moore, C. Unravelling the means to an end: RNA polymerase II transcription termination. *Nat Rev Mol Cell Biol*, 12(5):283–294, 2011. doi:10.1038/nrm3098. 12

Kumar-Sinha, C., Tomlins, S. A. and Chinnaiyan, A. M. Evidence of recurrent

gene fusions in common epithelial tumors. *Trends Mol Med*, 12(11):529–536, 2006. doi:10.1016/j.molmed.2006.09.005. 97

Kwak, H., Fuda, N. J., Core, L. J. and Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science*, 339(6122):950–953, 2013. doi:10.1126/science.1229386. 91

Lam, L. T., Pickeral, O. K., Peng, A. C., Rosenwald, A., Hurt, E. M., Giltnane, J. M., Averett, L. M., Zhao, H., Davis, R. E., Sathyamoorthy, M., Wahl, L. M., Harris, E. D., Mikovits, J. A., Monks, A. P., Hollingshead, M. G., Sausville, E. A. and Staudt, L. M. Genomic-scale measurement of mRNA turnover and the mechanisms of action of the anti-cancer drug flavopiridol. *Genome Biol*, 2(10):RESEARCH0041, 2001. 63

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009. doi:10.1186/gb-2009-10-3-r25. 21, 46

Lazo, P. A., DiPaolo, J. A. and Popescu, N. C. Amplification of the integrated viral transforming genes of human papillomavirus 18 and its 5'-flanking cellular sequence located near the myc protooncogene in HeLa cells. *Cancer Res*, 49(15):4305–4310, 1989. 50

Lefrançois, P., Euskirchen, G. M., Auerbach, R. K., Rozowsky, J., Gibson, T., Yellman, C. M., Gerstein, M. and Snyder, M. Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing. *BMC Genomics*, 10:37, 2009. doi:10.1186/1471-2164-10-37. 13

Lenhard, B., Sandelin, A. and Carninci, P. Metazoan promoters: emerging character-

istics and insights into transcriptional regulation. *Nat Rev Genet*, 13(4):233–245, 2012. doi:10.1038/nrg3163. 91, 92

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., Poh, H. M., Goh, Y., Lim, J., Zhang, J., Sim, H. S., Peh, S. Q., Mulawadi, F. H., Ong, C. T., Orlov, Y. L., Hong, S., Zhang, Z., Landt, S., Raha, D., Euskirchen, G., Wei, C.-L., Ge, W., Wang, H., Davis, C., Fisher-Aylor, K. I., Mortazavi, A., Gerstein, M., Gingeras, T., Wold, B., Sun, Y., Fullwood, M. J., Cheung, E., Liu, E., Sung, W.-K., Snyder, M. and Ruan, Y. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1-2):84–98, 2012. doi:10.1016/j.cell.2011.12.014. 103

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and , . G. P. D. P. S. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009. 21, 46

Lin, R.-K., Ho, C.-W., Liu, L. F. and Lyu, Y. L. Topoisomerase IIβ deficiency enhances camptothecin-induced apoptosis. *J Biol Chem*, 288(10):7182–7192, 2013. doi:10.1074/jbc.M112.415471. 116, 122

Listerman, I., Sapra, A. K. and Neugebauer, K. M. Cotranscriptional coupling of splicing factor recruitment and precursor messenger RNA splicing in mammalian cells. *Nat Struct Mol Biol*, 13(9):815–822, 2006. doi:10.1038/nsmb1135. 120

Liu, L. F. and Wang, J. C. Supercoiling of the DNA template during transcription. *Proc Natl Acad Sci U S A*, 84(20):7024–7027, 1987. 115

Liu, X., Bushnell, D. A., Silva, D.-A., Huang, X. and Kornberg, R. D. Initiation complex structure and promoter proofreading. *Science*, 333(6042):633–637, 2011. doi:10.1126/science.1206629. 8

Ljungman, M. and Hanawalt, P. C. Localized torsional tension in the DNA of human cells. *Proc Natl Acad Sci U S A*, 89(13):6055–6059, 1992. 143

Ljungman, M. and Hanawalt, P. C. Presence of negative torsional tension in the promoter region of the transcriptionally poised dihydrofolate reductase gene in vivo. *Nucleic Acids Res*, 23(10):1782–1789, 1995. 143

Ljungman, M. and Hanawalt, P. C. The anti-cancer drug camptothecin inhibits elongation but stimulates initiation of RNA polymerase II transcription. *Carcinogenesis*, 17(1):31–35, 1996. 18, 115, 116, 144

Ljungman, M. and Lane, D. P. Transcription - guarding the genome by sensing DNA damage. *Nat Rev Cancer*, 4(9):727–737, 2004. doi:10.1038/nrc1435. 116

Ljungman, M., O'Hagan, H. M. and Paulsen, M. T. Induction of ser15 and lys382 modifications of p53 by blockage of transcription elongation. *Oncogene*, 20(42):5964–5971, 2001. doi:10.1038/sj.onc.1204734. 116, 122

Ljungman, M. and Zhang, F. Blockage of RNA polymerase as a possible trigger for u.v. light-induced apoptosis. *Oncogene*, 13(4):823–831, 1996. 103, 117

Ljungman, M., Zhang, F., Chen, F., Rainbow, A. J. and McKay, B. C. Inhibition of RNA polymerase II as a trigger for the p53 response. *Oncogene*, 18(3):583–592, 1999. doi:10.1038/sj.onc.1202356. 73, 116, 117

Macville, M., Schröck, E., Padilla-Nash, H., Keck, C., Ghadimi, B. M., Zimonjic, D., Popescu, N. and Ried, T. Comprehensive and definitive molecular cytogenetic characterization of HeLa cells by spectral karyotyping. *Cancer Res*, 59(1):141–150, 1999. 50

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. and Hornik, K. *cluster: Cluster Analysis Basics and Extensions*, 2013. R package version 1.14.4 — For new features, see the 'Changelog' file (in the package source. 25, 105, 149

Marquardt, S., Escalante-Chong, R., Pho, N., Wang, J., Churchman, L. S., Springer, M. and Buratowski, S. A chromatin-based mechanism for limiting divergent noncoding transcription. *Cell*, 157(7):1712–1723, 2014. doi:10.1016/j.cell.2014.04.036. 14

Mason, P. B. and Struhl, K. Distinction and relationship between elongation rate and processivity of RNA polymerase II in vivo. *Mol Cell*, 17(6):831–840, 2005. doi:10.1016/j.molcel.2005.02.017. 142

McKay, B. C., Stubbert, L. J., Fowler, C. C., Smith, J. M., Cardamore, R. A. and Spronck, J. C. Regulation of ultraviolet light-induced gene expression by gene size. *Proc Natl Acad Sci U S A*, 101(17):6582–6586, 2004. doi:10.1073/pnas.0308181101. 123

Medzhitov, R. Origin and physiological roles of inflammation. *Nature*, 454(7203):428–435, 2008. doi:10.1038/nature07201. 63

Mercer, T. R., Neph, S., Dinger, M. E., Crawford, J., Smith, M. A., Shearwood, A.-M. J., Haugen, E., Bracken, C. P., Rackham, O., Stamatoyannopoulos, J. A., Filipovska, A. and Mattick, J. S. The human mitochondrial transcriptome. *Cell*, 146(4):645–658, 2011. doi:10.1016/j.cell.2011.06.051. 69

Micale, L., Muscarella, L. A., Marzulli, M., Augello, B., Tritto, P., D'Agruma, L., Zelante, L., Palumbo, G. and Merla, G. VHL frameshift mutation as target of nonsense-mediated mRNA decay in drosophila melanogaster and human HEK293 cell line. *J Biomed Biotechnol*, 2009:860761, 2009. doi:10.1155/2009/860761. 51

Min, J., Feng, Q., Li, Z., Zhang, Y. and Xu, R.-M. Structure of the catalytic domain of human DOT1L, a non-SET domain nucleosomal histone methyltransferase. *Cell*, 112(5):711–723, 2003. 144

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, 5(7):621–628, 2008. doi:10.1038/nmeth.1226. 21, 76, 83

Munchel, S. E., Shultzaberger, R. K., Takizawa, N. and Weis, K. Dynamic profiling of mRNA turnover reveals gene-specific and system-wide regulation of mRNA decay. *Mol Biol Cell*, 22(15):2787–2795, 2011. doi:10.1091/mbc.E11-01-0028. 63, 73

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349, 2008. doi:10.1126/science.1158441. 20

Ng, P., Wei, C.-L., Sung, W.-K., Chiu, K. P., Lipovich, L., Ang, C. C., Gupta, S., Shahab, A., Ridwan, A., Wong, C. H., Liu, E. T. and Ruan, Y. Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat Methods*, 2(2):105–111, 2005. doi:10.1038/nmeth733. 92

Nishioka, K., Rice, J. C., Sarma, K., Erdjument-Bromage, H., Werner, J., Wang, Y., Chuikov, S., Valenzuela, P., Tempst, P., Steward, R., Lis, J. T., Allis, C. D. and Reinberg, D. PR-Set7 is a nucleosome-specific methyltransferase that modifies lysine 20 of histone H4 and is associated with silent chromatin. *Mol Cell*, 9(6):1201–1213, 2002. 145

Nitiss, J. and Wang, J. C. Dna topoisomerase-targeting antitumor drugs can be studied in yeast. *Proc Natl Acad Sci U S A*, 85(20):7501–7505, 1988. 123

Nudler, E., Mustaev, A., Lukhtanov, E. and Goldfarb, A. The RNA-DNA hybrid maintains the register of transcription by preventing backtracking of RNA polymerase. *Cell*, 89(1):33–41, 1997. 11

O'Connell, R. M., Taganov, K. D., Boldin, M. P., Cheng, G. and Baltimore, D. MicroRNA-155 is induced during the macrophage inflammatory response. *Proc Natl Acad Sci U S A*, 104(5):1604–1609, 2007. doi:10.1073/pnas.0610731104. 72

Ohtsu, M., Kawate, M., Fukuoka, M., Gunji, W., Hanaoka, F., Utsugi, T., Onoda, F. and Murakami, Y. Novel DNA microarray system for analysis of nascent mRNAs. *DNA Res*, 15(4):241–251, 2008. doi:10.1093/dnares/dsn015. 15, 29, 63, 64

Okada, Y., Feng, Q., Lin, Y., Jiang, Q., Li, Y., Coffield, V. M., Su, L., Xu, G. and Zhang, Y. hDOT1L links histone methylation to leukemogenesis. *Cell*, 121(2):167–178, 2005. doi:10.1016/j.cell.2005.02.020. 145

Onder, T. T., Kara, N., Cherry, A., Sinha, A. U., Zhu, N., Bernt, K. M., Cahan, P., Marcarci, B. O., Unternaehrer, J., Gupta, P. B., Lander, E. S., Armstrong, S. A. and Daley, G. Q. Chromatin-modifying enzymes as modulators of reprogramming. *Nature*, 483(7391):598–602, 2012. doi:10.1038/nature10953. 144

Pal, M. and Luse, D. S. The initiation-elongation transition: lateral mobility of RNA in RNA polymerase II complexes is greatly reduced at +8/+9 and absent by +23. *Proc Natl Acad Sci U S A*, 100(10):5700–5705, 2003. doi:10.1073/pnas.1037057100. 8

Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*, 10(10):669–680, 2009. doi:10.1038/nrg2641. 22

Paulsen, M. T., Veloso, A., Prasad, J., Bedi, K., Ljungman, E. A., Magnuson, B.,

Wilson, T. E. and Ljungman, M. Use of Bru-Seq and BruChase-seq for genome-wide assessment of the synthesis and stability of RNA. *Methods*, 2013a. doi: 10.1016/j.ymeth.2013.08.015. 15, 94, 103, 104, 106, 135, 145, 146

Paulsen, M. T., Veloso, A., Prasad, J., Bedi, K., Ljungman, E. A., Tsan, Y.-C., Chang, C.-W., Tarrier, B., Washburn, J. G., Lyons, R., Robinson, D. R., Kumar-Sinha, C., Wilson, T. E. and Ljungman, M. Coordinated regulation of synthesis and stability of RNA during the acute TNF-induced proinflammatory response. *Proc Natl Acad Sci U S A*, 110(6):2240–2245, 2013b. doi:10.1073/pnas.1219192110. 15, 30, 45, 54, 91, 94, 99, 101, 103, 104, 106, 112, 117, 118, 120, 135, 145, 146

Pereira, F. J. C., do Céu Silva, M., Picano, I., Seixas, M. T., Ferrão, A., Faustino, P. and Romo, L. Human alpha2-globin nonsense-mediated mRNA decay induced by a novel alpha-thalassaemia frameshift mutation at codon 22. *Br J Haematol*, 133(1):98–102, 2006. doi:10.1111/j.1365-2141.2006.05971.x. 51

Peterlin, B. M. and Price, D. H. Controlling the elongation phase of transcription with P-TEFb. *Mol Cell*, 23(3):297–305, 2006. doi:10.1016/j.molcel.2006.06.014. 10

Plo, I., Liao, Z. Y., Barceló, J. M., Kohlhagen, G., Caldecott, K. W., Weinfeld, M. and Pommier, Y. Association of XRCC1 and tyrosyl DNA phosphodiesterase (Tdp1) for the repair of topoisomerase I-mediated DNA lesions. *DNA Repair (Amst)*, 2(10):1087–1100, 2003. 115

Pommier, Y. Topoisomerase I inhibitors: camptothecins and beyond. *Nat Rev Cancer*, 6(10):789–802, 2006. doi:10.1038/nrc1977. 19, 115, 120

Pommier, Y. Drugging topoisomerases: lessons and challenges. *ACS Chem Biol*, 8(1):82–95, 2013. doi:10.1021/cb300648v. 115, 123, 125

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M. S., Mapendano, C. K., Schierup, M. H. and Jensen, T. H. RNA exosome depletion reveals transcription upstream of active human promoters. *Science*, 322(5909):1851–1854, 2008. doi:10.1126/science.1164096. 66, 120

Quinlan, A. R. and Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010. doi:10.1093/bioinformatics/btq033. 21, 46, 81, 82

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0. 21, 46

Rabani, M., Levin, J. Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., Amit, I. and Regev, A. Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat Biotechnol*, 29(5):436–442, 2011. doi:10.1038/nbt.1861. 15, 22, 29, 63, 167

Raghavan, A., Ogilvie, R. L., Reilly, C., Abelson, M. L., Raghavan, S., Vasdewani, J., Krathwohl, M. and Bohjanen, P. R. Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic Acids Res*, 30(24):5529–5538, 2002. 63

Ram, O., Goren, A., Amit, I., Shoresh, N., Yosef, N., Ernst, J., Kellis, M., Gymrek, M., Issner, R., Coyne, M., Durham, T., Zhang, X., Donaghey, J., Epstein, C. B., Regev, A. and Bernstein, B. E. Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell*, 147(7):1628–1639, 2011. doi:10.1016/j.cell.2011.09.057. 153

Rao, B., Shibata, Y., Strahl, B. D. and Lieb, J. D. Dimethylation of histone H3 at lysine 36 demarcates regulatory and nonregulatory chromatin genome-wide. *Mol Cell Biol*, 25(21):9447–9459, 2005. doi:10.1128/MCB.25.21.9447-9459.2005. 142

Rasmussen, E. B. and Lis, J. T. In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. *Proc Natl Acad Sci U S A*, 90(17):7923–7927, 1993. 8

Reeve, J. N. Archaeal chromatin and transcription. *Mol Microbiol*, 48(3):587–598, 2003. 7

Roberts, G. C., Gooding, C., Mak, H. Y., Proudfoot, N. J. and Smith, C. W. Co-transcriptional commitment to alternative splice site selection. *Nucleic Acids Res*, 26(24):5568–5572, 1998. 11

Rockx, D. A., Mason, R., van Hoffen, A., Barton, M. C., Citterio, E., Bregman, D. B., van Zeeland, A. A., Vrieling, H. and Mullenders, L. H. UV-induced inhibition of transcription involves repression of transcription initiation and phosphorylation of RNA polymerase II. *Proc Natl Acad Sci U S A*, 97(19):10503–10508, 2000. doi: 10.1073/pnas.180169797. 102

Rodriguez, J., Menet, J. S. and Rosbash, M. Nascent-seq indicates widespread co-transcriptional RNA editing in Drosophila. *Mol Cell*, 47(1):27–37, 2012. doi: 10.1016/j.molcel.2012.05.002. 15

Rowley, J. D. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature*, 243(5405):290–293, 1973. 97

Sakai, A., Sakasai, R., Kakeji, Y., Kitao, H. and Maehara, Y. PARP and CSB

modulate the processing of transcription-mediated DNA strand breaks. *Genes Genet Syst*, 87(4):265–272, 2012. 116, 122

Sakasai, R., Teraoka, H., Takagi, M. and Tibbetts, R. S. Transcription-dependent activation of ataxia telangiectasia mutated prevents DNA-dependent protein kinase-mediated cell death in response to topoisomerase I poison. *J Biol Chem*, 285(20):15201–15208, 2010. doi:10.1074/jbc.M110.101808. 116

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y. and Hume, D. A. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet*, 8(6):424–436, 2007. doi:10.1038/nrg2026. 91

Sanyal, A., Lajoie, B. R., Jain, G. and Dekker, J. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113, 2012. doi:10.1038/nature11279. 91

Sauerbier, W. and Hercules, K. Gene and transcription unit mapping by radiation effects. *Annu Rev Genet*, 12:329–363, 1978. doi:10.1146/annurev.ge.12.120178.001553. 93, 102

Saunders, A., Core, L. J. and Lis, J. T. Breaking barriers to transcription elongation. *Nat Rev Mol Cell Biol*, 7(8):557–567, 2006. doi:10.1038/nrm1981. 8

Schwalb, B., Schulz, D., Sun, M., Zacher, B., Dümcke, S., Martin, D. E., Cramer, P. and Tresch, A. Measurement of genome-wide RNA synthesis and decay rates with Dynamic Transcriptome Analysis (DTA). *Bioinformatics*, 28(6):884–885, 2012. doi:10.1093/bioinformatics/bts052. 29

Schwanhusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen,

W. and Selbach, M. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, 2011. doi:10.1038/nature10098. 29, 63

Selby, C. P. and Sancar, A. Cockayne syndrome group B protein enhances elongation by RNA polymerase II. *Proc Natl Acad Sci U S A*, 94(21):11205–11209, 1997. 137, 143

Seoighe, C. and Korir, P. K. Evidence for intron length conservation in a set of mammalian genes associated with embryonic development. *BMC Bioinformatics*, 12 Suppl 9:S16, 2011. doi:10.1186/1471-2105-12-S9-S16. 133

Shandilya, J. and Roberts, S. G. E. The transcription cycle in eukaryotes: from productive initiation to RNA polymerase II recycling. *Biochim Biophys Acta*, 1819(5):391–400, 2012. doi:10.1016/j.bbagrm.2012.01.010. 7, 19, 133

Shi, Y., Di Giammartino, D. C., Taylor, D., Sarkeshik, A., Rice, W. J., Yates, 3rd, J. R., Frank, J. and Manley, J. L. Molecular architecture of the human pre-mRNA 3' processing complex. *Mol Cell*, 33(3):365–376, 2009. doi:10.1016/j.molcel.2008.12.028. 11

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., Kodzius, R., Watahiki, A., Nakamura, M., Arakawa, T., Fukuda, S., Sasaki, D., Podhajska, A., Harbers, M., Kawai, J., Carninci, P. and Hayashizaki, Y. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A*, 100(26):15776–15781, 2003. doi:10.1073/pnas.2136655100. 92

Shukla, S. and Oberdoerffer, S. Co-transcriptional regulation of alternative pre-mRNA splicing. *Biochim Biophys Acta*, 1819(7):673–683, 2012. doi:10.1016/j.bbagrm.2012.01.014. 11, 133, 143

Singh, J. and Padgett, R. A. Rates of in situ transcription and splicing in large human genes. *Nat Struct Mol Biol*, 16(11):1128–1133, 2009. doi:10.1038/nsmb.1666. 20, 24, 134, 137, 142

Smolle, M. and Workman, J. L. Transcription-associated histone modifications and cryptic transcription. *Biochim Biophys Acta*, 1829(1):84–97, 2013. doi:10.1016/j.bbagrm.2012.08.008. 142

Smyth, G. K., Michaud, J. and Scott, H. S. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, 21(9):2067–2075, 2005. doi:10.1093/bioinformatics/bti270. 148

Sordet, O., Redon, C. E., Guirouilh-Barbat, J., Smith, S., Solier, S., Douarre, C., Conti, C., Nakamura, A. J., Das, B. B., Nicolas, E., Kohn, K. W., Bonner, W. M. and Pommier, Y. Ataxia telangiectasia mutated activation by transcription- and topoisomerase I-induced DNA double-strand breaks. *EMBO Rep*, 10(8):887–893, 2009. doi:10.1038/embor.2009.97. 116

Spieth, J., Brooke, G., Kuersten, S., Lea, K. and Blumenthal, T. Operons in C. elegans: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. *Cell*, 73(3):521–532, 1993. 96

Spitz, F. and Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet*, 13(9):613–626, 2012. doi:10.1038/nrg3207. 91, 133

Squires, S., Ryan, A. J., Strutt, H. L. and Johnson, R. T. Hypersensitivity of Cockayne's syndrome cells to camptothecin is associated with the generation of abnormally high levels of double strand breaks in nascent DNA. *Cancer Res*, 53(9):2012–2019, 1993. 116, 121, 122, 124

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, 2005. doi:10.1073/pnas.0506580102. 138, 149

Sun, M., Schwalb, B., Schulz, D., Pirkl, N., Etzold, S., Larivire, L., Maier, K. C., Seizl, M., Tresch, A. and Cramer, P. Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome Res*, 22(7):1350–1359, 2012. doi:10.1101/gr.130161.111. 16, 29

Sutherland, H. and Bickmore, W. A. Transcription factories: gene expression in unions? *Nat Rev Genet*, 10(7):457–466, 2009. doi:10.1038/nrg2592. 133

Swinburne, I. A. and Silver, P. A. Intron delays and transcriptional timing during development. *Dev Cell*, 14(3):324–330, 2008. doi:10.1016/j.devcel.2008.02.002. 133, 145, 166

Taganov, K. D., Boldin, M. P., Chang, K.-J. and Baltimore, D. NF-kappaB-dependent induction of microRNA miR-146, an inhibitor targeted to signaling proteins of innate immune responses. *Proc Natl Acad Sci U S A*, 103(33):12481–12486, 2006. doi:10.1073/pnas.0605298103. 72

Takashima, Y., Ohtsuka, T., Gonzlez, A., Miyachi, H. and Kageyama, R. Intronic delay is essential for oscillatory expression in the segmentation clock. *Proc Natl Acad Sci U S A*, 108(8):3300–3305, 2011. doi:10.1073/pnas.1014418108. 133

Tani, H., Mizutani, R., Salam, K. A., Tano, K., Ijiri, K., Wakamatsu, A., Isogai, T., Suzuki, Y. and Akimitsu, N. Genome-wide determination of RNA stability

reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Res*, 22(5):947–956, 2012. doi:10.1101/gr.130559.111. 16, 22, 29, 63, 64, 68

Tardat, M., Brustel, J., Kirsh, O., Lefevbre, C., Callanan, M., Sardet, C. and Julien, E. The histone H4 Lys 20 methyltransferase PR-Set7 regulates replication origins in mammalian cells. *Nat Cell Biol*, 12(11):1086–1093, 2010. doi:10.1038/ncb2113. 145

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E. and Stamatoyannopoulos, J. A. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 2012. doi:10.1038/nature11232. 103

Tian, B., Nowak, D. E. and Brasier, A. R. A TNF-induced gene expression program under oscillatory NF-kappaB control. *BMC Genomics*, 6:137, 2005. doi:10.1186/1471-2164-6-137. 17, 63, 70

Tölg, C., Hofmann, M., Herrlich, P. and Ponta, H. Splicing choice from ten variant exons establishes CD44 variability. *Nucleic Acids Res*, 21(5):1225–1229, 1993. 52

Tornaletti, S. and Hanawalt, P. C. Effect of DNA lesions on transcription elongation. *Biochimie*, 81(1-2):139–146, 1999. 18, 93

Trapnell, C., Pachter, L. and Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009. doi:10.1093/bioinformatics/btp120. 21, 46, 81

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. and Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–515, 2010. doi:10.1038/nbt.1621. 23

Tripathi, V., Ellis, J. D., Shen, Z., Song, D. Y., Pan, Q., Watt, A. T., Freier, S. M., Bennett, C. F., Sharma, A., Bubulya, P. A., Blencowe, B. J., Prasanth, S. G. and Prasanth, K. V. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell*, 39(6):925–938, 2010. doi:10.1016/j.molcel.2010.08.011. 66

Valen, E., Pascarella, G., Chalk, A., Maeda, N., Kojima, M., Kawazu, C., Murata, M., Nishiyori, H., Lazarevic, D., Motti, D., Marstrand, T. T., Tang, M.-H. E., Zhao, X., Krogh, A., Winther, O., Arakawa, T., Kawai, J., Wells, C., Daub, C., Harbers, M., Hayashizaki, Y., Gustincich, S., Sandelin, A. and Carninci, P. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res*, 19(2):255–265, 2009. doi:10.1101/gr.084541.108. 92

Veloso, A., Biewen, B., Paulsen, M. T., Berg, N., Carmo de Andrade Lima, L., Prasad, J., Bedi, K., Magnuson, B., Wilson, T. E. and Ljungman, M. Genome-wide transcriptional effects of the anti-cancer agent camptothecin. *PLoS One*, 8(10):e78190, 2013. doi:10.1371/journal.pone.0078190. 103, 144, 145

Veloso, A., Kirkconnell, K. S., Magnuson, B., Biewen, B., Paulsen, M. T., Wilson, T. E. and Ljungman, M. Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res*, 24(6):896–905, 2014. doi:10.1101/gr.171405.113. 103, 105

Venters, B. J. and Pugh, B. F. Genomic organization of human transcription initiation complexes. *Nature*, 502(7469):53–58, 2013. doi:10.1038/nature12535. 91

Wada, T., Takagi, T., Yamaguchi, Y., Ferdous, A., Imai, T., Hirose, S., Sugimoto, S., Yano, K., Hartzog, G. A., Winston, F., Buratowski, S. and Handa, H. DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes Dev*, 12(3):343–356, 1998a. 9, 19

Wada, T., Takagi, T., Yamaguchi, Y., Watanabe, D. and Handa, H. Evidence that P-TEFb alleviates the negative effect of DSIF on RNA polymerase II-dependent transcription in vitro. *EMBO J*, 17(24):7395–7403, 1998b. doi:10.1093/emboj/17.24.7395. 9

Wada, Y., Ohta, Y., Xu, M., Tsutsumi, S., Minami, T., Inoue, K., Komura, D., Kitakami, J., Oshida, N., Papantonis, A., Izumi, A., Kobayashi, M., Meguro, H., Kanki, Y., Mimura, I., Yamamoto, K., Mataki, C., Hamakubo, T., Shirahige, K., Aburatani, H., Kimura, H., Kodama, T., Cook, P. R. and Ihara, S. A wave of nascent transcription on activated human genes. *Proc Natl Acad Sci U S A*, 106(43):18357–18361, 2009. doi:10.1073/pnas.0902573106. 24, 134, 142

Wang, J. C. Cellular roles of DNA topoisomerases: a molecular perspective. *Nat Rev Mol Cell Biol*, 3(6):430–440, 2002. doi:10.1038/nrm831. 18

Wang, K. C. and Chang, H. Y. Molecular mechanisms of long noncoding RNAs. *Mol Cell*, 43(6):904–914, 2011. doi:10.1016/j.molcel.2011.08.018. 53

Wang, Y., Fairley, J. A. and Roberts, S. G. E. Phosphorylation of TFIIB links transcription initiation and termination. *Curr Biol*, 20(6):548–553, 2010. doi:10.1016/j.cub.2010.01.052. 9

Wang, Z., Gerstein, M. and Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, 2009. doi:10.1038/nrg2484. 22

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I. and Young, R. A. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–319, 2013. doi:10.1016/j.cell.2013.03.035. 100

Wu, J. and Liu, L. F. Processing of topoisomerase I cleavable complexes into DNA damage by transcription. *Nucleic Acids Res*, 25(21):4181–4186, 1997. 115, 116, 121

Wuarin, J. and Schibler, U. Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. *Mol Cell Biol*, 14(11):7219–7225, 1994. 14

Yamaguchi, Y., Takagi, T., Wada, T., Yano, K., Furuya, A., Sugimoto, S., Hasegawa, J. and Handa, H. NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell*, 97(1):41–51, 1999. 9, 19

Zentner, G. E., Saiakhova, A., Manaenkov, P., Adams, M. D. and Scacheri, P. C. Integrative genomic analysis of human ribosomal DNA. *Nucleic Acids Res*, 39(12):4949–4960, 2011. doi:10.1093/nar/gkq1326. 69

Zentner, G. E. and Scacheri, P. C. The chromatin fingerprint of gene enhancer elements. *J Biol Chem*, 287(37):30888–30896, 2012. doi:10.1074/jbc.R111.296491. 92

Zhang, Y.-W., Regairaz, M., Seiler, J. A., Agama, K. K., Doroshow, J. H. and Pommier, Y. Poly(ADP-ribose) polymerase and XPF-ERCC1 participate in distinct pathways for the repair of topoisomerase I-induced DNA damage in mammalian cells. *Nucleic Acids Res*, 39(9):3607–3620, 2011. doi:10.1093/nar/gkq1304. 125

Zhu, Y., Pe'ery, T., Peng, J., Ramanathan, Y., Marshall, N., Marshall, T., Amendt, B., Mathews, M. B. and Price, D. H. Transcription elongation factor P-TEFb is required for HIV-1 tat transactivation in vitro. *Genes Dev*, 11(20):2622–2632, 1997. 19