Capturing transcriptional dynamics using nascent RNA sequencing

by

Killeen S. Kirkconnell

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Human Genetics)
in the University of Michigan
2016

Doctoral Committee:

Professor Mats Ljungman, Co-Chair
Associate Professor Thomas E. Wilson, Co-Chair
Professor Thomas W. Glover
Assistant Professor Ryan E. Mills
Assistant Professor Indika Rajapakse

ACKNOWLEDGEMENTS

Thank you to my mentors, my labmates, and my thesis committee members for their contributions to the following work. Thank you to my family and friends for their continued support.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Transcription plays an essential role in establishing cellular identity and functional control, and each step of the transcriptional process represents a potential point of regulation. Methods that survey RNA abundance within a given cell type can be used to study steady state levels based on the balance between RNA synthesis and turnover. However, a thorough examination of how the transcriptional process contributes to gene expression regulation requires observation of the earliest stages of RNA production. The work presented in this dissertation explored early transcriptional events using Bru-seq nascent RNA sequencing. This technique allows for the detection of changes in RNA synthesis across gene bodies prior to the completion of full length transcripts. Using Bru-seq, I was able to monitor genome-wide expression changes during the first two hours following serum stimulation of human fibroblast cells. This led to the identification of over 2000 genes that demonstrated a transcriptional response to serum activation, including a novel group of genes which were immediately repressed. Response genes were categorized according to distinct transcriptional induction and repression patterns, providing new candidate gene groups for studying common regulatory mechanisms during global transcriptional responses. Additionally, I took advantage of the Bru-seq technique to follow transcription elongation in long genes over time. The dataset revealed how gene length influences transcriptional timing, and demonstrated that a set of genes with different sizes can be simultaneously induced but expressed at various times. Because relative gene size is conserved in mammals, this suggests that gene length plays an important role in maintaining proper temporal expression patterns. Lastly, I used a modification of this technique, BruUV-seq, to identify active enhancer elements based on enhancer RNA production, and to observe the effects of inhibiting BRD4, an important transcriptional coactivator and enhancer chromatin regulator. This dataset indicated that BRD4 inhibition results in immediate disruption of enhancer transcription, and that BRD4 function is required for the maintenance of enhancer activity. Collectively, this work extends our knowledge of how early transcriptional events impact the regulation of

gene expression, and provides a foundation for future studies exploring the precise mechanisms which determine cellular identity and functional control.

CHAPTER I

Introduction

1.1 Transcription

Gene expression is a fundamental process that allows genetic information to be converted into a functional product. All living organisms utilize the process of gene expression to generate the molecular machinery required for cellular activity. The human genome contains the genetic blueprints for approximately 20,000 different genes[1]. While each cell possesses a complete set of every single gene, distinct cellular functions are established through expression of only a subset of these genes. This precise regulation of gene expression allows for hundreds of different specialized cell types[2]. During the flow of information from gene to product, transcription is a critical first step in which DNA is used to synthesize RNA transcripts. Each stage of transcription acts as a point of regulation by which cells can adjust the production levels of gene products and thereby modulate their functional capabilities. Therefore, transcription is essential for establishing cellular identity and allowing cells to respond to environmental stimuli. A comprehensive examination of each stage of transcription is important for a deeper understanding of the complex mechanisms which contribute to gene expression regulation. The following review of the different stages in the transcriptional process focuses on eukaryotic transcription, namely in metazoans.

1.1.1 Enhancers

Transcription is controlled by an intricate network of enhancer regulatory elements which dictate precise spatiotemporal patterns of gene expression. Enhancers are genetic elements that contain transcription factor binding motifs and enable increased transcription levels[3]. They are able to regulate the transcription of genes that are distantly located on the linear chromosome. Much of the regulation dictating cell type-specific gene expression is driven by enhancer activity.

Enhancers can exist in various states, which are often identified based on histone marks and transcription factor binding signatures[4-6]. These different states represent the transition from enhancer selection to activation. Inactive enhancers are obscured by compact or closed chromatin and do not display any of the characteristic enhancer chromatin marks or transcription factor binding. Primed enhancers are within regions of open chromatin and bound by transcription factors[7,8]. This occurs via pioneer transcription factors (TFs) which bind and facilitate nucleosome remodeling[9], allowing lineage-determining TFs to selectively bind cell type-specific enhancers[10]. This can lead to subsequent recruitment of collaborative TFs and transcriptional cofactors, such as histone methyltransferases. Poised enhancers are enriched for histone H3 lysine 4 monomethylation (H3K4me1) and depletion of H3K4me3 compared with promoters[11,12]. Activation of enhancers occurs through the recruitment of additional TFs and cofactors, including histone acetyltransferases. Active enhancers are associated with H3K27 acetylation (H3K27ac)[13]. Signal dependent TFs bind to enhancers in response to cellular cues and can activate both common genes sets or cell type-specific gene sets[14]. Binding of RNA polymerase II (RNAPII) can result in the initiation of transcription, which is often bidirectional, at the enhancer[15,16]. Additional cofactors can bind to the acetylated histone tail, such as bromodomain-containing protein 4 (BRD4), which likely promotes elongation of enhancer RNA (eRNA) transcripts[17]. Chromatin looping factors facilitate interactions between enhancers and promoters, however the mechanisms behind these conformational changes are largely unknown[18]. The large Mediator complex, which can facilitate binding interactions between itself, RNAPII, and TFs, is critical for chromatin organization and looping[19]. These physical interactions between the two regions may allow transcriptional machinery to be loaded at the enhancer and then transferred to the promoter[20,21]. Promoters may be able to transfer transcriptional factors over to enhancers as well[16]. It is likely that the collaboration between DNA binding factors at both promoters and enhancers allows for maximum transcriptional output[20].

Enhancers are commonly transcribed and produce eRNA[22,23]. Most eRNAs tend to be short, non-spliced, and non-polyadenylated transcripts which are unstable and rapidly degraded by the exosome[15,16,23]. These transcript characteristics are similar to those of promoter upstream transcripts (PROMPTs). While eRNAs could be considered a type of long non-coding RNA (lncRNA) based on their length, many eRNAs are not represented in lncRNA databases, which may be related to their unstable nature[24]. Changes in eRNA production are correlated with changes in the expression of nearby genes[15,16]. Therefore, it is thought that eRNA transcription is a defining mark of enhancer activity. However, it is possible that enhancers that do not show evidence of transcription are also functional[23].

The role of eRNA production in enhancer activation and the regulation of target gene expression is not currently understood. There are several current non-exclusive models for the function of enhancer transcription[6]. One model hypothesizes that eRNAs may not be functional and merely represent transcriptional noise resulting from the accumulation of transcriptional machinery. An alternative model postulates that it is not eRNA itself, but rather the act of transcribing eRNAs that is important for enhancer activity[25,26]. Enhancer transcription may be important for maintaining open chromatin and histone modifications[27]. Thirdly, eRNAs themselves may play functional roles in gene activation. It has been suggested that eRNAs may be important for facilitating enhancer-promoter looping interactions[28-30], RNAPII loading[31], promoting pause release and productive elongation[32], the recruitment of transcription factors or cofactors[28,30], and the binding and inhibition of transcriptional repressors[32].

Large, enhancer-dense regions have been described as super-enhancers[33-35]. These regions tend to be thousands of basepairs in length compared to typical enhancers which are hundreds of basepairs long. There are approximately 300-500 super-enhancers in a given cell, and they are often cell-type specific and adjacent to genes that are key drivers of cellular identity[33,36,37]. They also produce high levels of eRNAs[38,39]. The original method to define super-enhancers in mouse embryonic stem cells (mESCs) employed three steps[36]. First, they identified regions bound by master transcription factors important for maintaining pluripotency. Second, they stitched together regions which were within a 12.5 kb range. Lastly, they selected a subset of these regions which had the high densities of Mediator complex binding. This was done by ranking all enhancers and selecting a cutoff based on binding signal. Several other studies have used similar methods to define super-enhancer regions in other cell types. Typically, different factors are used for the first and last steps, often lineage specific master regulators or H3K27ac levels[34].

Stretch enhancers is a term that has been used to describe large non-stitched regions (>3kb) containing high levels of certain chromatin marks such as H3K27ac[40]. These regions are also shown to be cell type-specific and contain transcription factor binding motifs that play a role in cellular identity. While the number of super-enhancer regions in a given cell line tends to be in the hundreds, the number of stretch enhancer regions tends to be in the thousands[35]. For a given cell line, it has been found that most super enhancers overlap stretch enhancer regions, and so it seems that super-enhancers are a subset of stretch enhancers[5,40].

The function of super-enhancers requires additional investigation. Most of the studies describing super-enhancers do not test functionality. A recent study used luciferase reporter vectors to test interactions between the active enhancers within a super-enhancer cluster[41]. They found that the elements were not additive or synergistic, but were all necessary for optimal transcriptional activity. They also used genome editing to delete individual enhancer elements, which led to decreases in expression of nearby associated genes[41]. While many super-enhancers seem to be made up of multiple enhancer components, some of the defined super-enhancers in mESCs consisted of a single enhancer[36]. Also, it seems that most stitched enhancers are not super-enhancers, so clustering of nearby enhancers is not sufficient to identify super-enhancers[34,36]. In order to further characterize the function of super-enhancers in gene expression, it may be necessary to first clearly define what constitutes a super-enhancer and develop methods to reliably identify them.

1.1.2 Initiation

Transcription begins with initiation, which has long been recognized as a key regulation point during gene expression. Initiation involves the binding of RNAPII at promoter sequences to begin RNA synthesis. Gene promoters are the sequence elements that act as binding sites for the transcriptional machinery. Promoters act to appropriately align the machinery in order to instruct the direction of transcription. Elements of core promoters include the TATA element, B recognition element (BRE), initiator (Inr), and downstream promoter element (DPE)[42]. Promoters may include one or a combination of these elements, but none of them are essential for promoter function.

RNAPII is the enzyme responsible for the transcription of protein-coding genes as well as the production of some non-coding RNAs such as microRNAs (miRNAs), small nuclear RNAs (snRNAs), and eRNAs[15,43-45]. The core RNAPII enzyme is complex and consists of 12 subunits[43]. In order for initiation to occur, RNAPII binds to promoter DNA along with a group of general transcription factors to form the pre-initiation complex (PIC)[43,46,47]. The general transcription factors (GTFs) include TFIIB, TFIID, TFIIE, TFIIF, and TFIIH. TFIID contains the TATA box-binding protein (TBP), which is required for transcription at all promoters. TBP binding to the TATA element induces a bend in the DNA[48]. TFIID also contains several TBP-associated factors (TAFs) which have promoter-specific functions. TAFs have been shown to bind to Inr and DPE promoter elements[49-51]. Subsequent TFIIA and TFIIB binding around TFIID act to stabilize the complex on the DNA. RNAPII and TFIIF are then recruited to the complex, followed by TFIIE and TFIIH. When the PIC is formed in the presence of nucleoside triphosphates, it induces a conformational change in the DNA, resulting in promoter melting and formation of the transcription bubble. Initiation occurs when the DNA template strand passes near the RNAPII active sight and synthesis of RNA begins. Frequently, short abortive RNA products are made before RNAPII transitions into productive initiation[52]. Once the RNA-DNA hybrid reaches a certain length, subsequent synthesis results in the newly made RNA separating from the DNA template and entering the RNA exit channel of RNAPII[53]. RNAPII can then undergo promoter clearance and release its interactions with the GTFs[54]. During this process, TFIIH phosphorylates serine 5 (Ser5) residues of the RNAPII carboxy-terminal domain (CTD). RNAPII can then associate with elongation factors and begin to elongate the RNA transcript.

1.1.3 Elongation

Elongation is also now recognized as an important regulatory step in gene expression. During early elongation, RNA synthesis can pause and accumulate in the region that is 30-60 nucleotides downstream of the transcription start site (TSS), and this is called promoter-proximal pausing of RNAPII[55,56]. Pausing factors, such as negative elongation factor (NELF)[57] and DRB-sensitivity inducing factor (DSIF)[58], associate with and stabilize RNAPII. Signaling events can result in the recruitment of TFs, such as MYC[59] or NF-κB[60], that recruit positive transcription factor-b (P-TEFb) directly or through cofactors such as BRD4[61,62] and the super elongation complex (SEC)[63]. Activation of P-TEFb results in the phosphorylation of NELF, causing its disassociation, and the phosphorylation of DSIF, which then becomes a positive elongation factor[64]. P-TEFb also phosphorylates serine 2 (Ser2) on the CTD of RNAPII, which acts as a scaffold for RNA-processing factors and chromatin modifiers that play a role in transcription elongation[64]. This leads to pause release and the transition into productive elongation. Pausing can function to maintain open chromatin at the promoters of primed and active genes so that they are accessible to TFs[65]. While pause release appears to be a necessary step during transcription, the accumulation of paused polymerases downstream of the promoter region only occurs in subset of genes in a cell type and treatment specific manner[56]. Therefore, pause release acts as a potential regulatory step for modulating gene expression at active genes.

Once RNAPII is released from the promoter pausing region and productive elongation begins, RNA synthesis takes place across the gene body[66]. The rate of transcription elongation was measured genome-wide by our lab and others, and was seen to be variable among genes and across cell types[67-71]. Elongation rate also appears to vary across the gene body, with rates of around 0.5kb/min within the first few kilobases, and 2-5kb/min after around 15kb. Changes in the composition of the transcription machinery as it travels across longer genes may be responsible for this acceleration of elongation[56]. There are also certain gene features which reduce the speed of the elongating RNAPII. These include exons and the mRNA cleavage and polyadenylation sites, and may be related to co-transcriptional RNA processing and transcription termination[56]. Additionally, epigenetic marks have been shown to be correlated with elongation rates. Therefore, while there are fixed gene features that influence elongation rate, cells can also potentially fine tune elongation rates through chromatin modifications.

1.1.4 Termination

Termination is the end step in the transcriptional process when nascent RNA and RNAPII are released from the DNA template[72]. Cleavage and polyadenylation specificity factor (CPSF) binds to RNAPII while cleavage stimulatory factor (CstF) and cleavage factors I and II (CFI and CFII) bind to the Ser2-phosphorylated CTD. The polyadenylation signal (PAS) in the 3' untranslated region (UTR) of the nascent RNA is recognized by CPSF and CstF. Interactions between CPSF, RNAPII, and the PAS are thought to result in pausing of RNAPII, which may facilitate termination events. The RNA is cleaved 18-30 nucleotides downstream of the PAS. Senataxin (SETX) has been proposed to resolve R-loops at the 3'end fragment of the cleaved RNA[73]. This allows the 5'-3' exoribonuclease XRN2 to access the end fragment for degradation. The mechanism by which the elongating RNAPII complex dissociates from the template during termination is unclear. One model, the torpedo model, proposes that XRN2 degradation of the end fragment acts to release RNAPII from the template[74]. An alternative allosteric model suggests that transcription of the PAS results in a conformational change in RNAPII which destabilizes the elongation complex[75].

Transcription termination is also viewed as a step that regulates gene expression and pervasive transcription[72]. Several instances of gene regulation through transcription termination have been observed in yeast, and these mechanisms may play regulatory roles in metazoans as well. Through autoregulatory negative feedback loops, some genes have been shown to promote early termination of transcription and RNA degradation when protein levels are high[76]. Alternatively, genes can utilize regulated attenuation to normally activate early termination, but then allow full transcription under special conditions[77]. Termination is also thought to regulate bidirectional transcription at promoters and target PROMPTs for degradation[78]. PASs are depleted in the direction of the mRNA and enriched in the divergent orientation. Small nuclear ribonucleoproteins (snRNPs) bind at sites near infrequent PASs on the coding side, thereby blocking recognition by the CPSF-CF complex and allowing transcription elongation to continue. On the divergent side, frequent PASs are recognized by the CPSF-CF complex and the cap-binding complex (CBC)-ARS2, which promote early termination and exosome targeting of RNA.

1.1.5 RNA processing and stability

Before protein coding RNAs undergo translation they are subject to several RNA processing events. These events are tightly coupled with each other and transcription, and contribute to the diversity of RNA transcripts. Therefore, RNA processing also plays a role in the regulation of gene expression.

Capping of the 5'end of nascent RNA is the first processing step. Addition of the 7-methylguanosine cap occurs after transcription of the first 25-30 nucleotides and requires the activities of an RNA triphosphatase, guanylyltransferase, and methyltransferase[79]. The capping enzyme binds to the Ser5 phosphorylated CTD of the elongating RNAPII[80]. Capping is important for protecting the transcript from degradation by 5'-3' exonucleases as well as initiating translation through interactions with the ribosome[81].

Introns within the pre-mRNA transcript must be removed through splicing. Splicing is catalyzed by the spliceosome, which is comprised of five snRNPs and a large number of protein components[82]. Core sequence elements include the 5' and 3' splice sites, branch point, and polypyrimidine tract, which are all bound by spliceosome components[81]. Splicing is thought to occur cotranscriptionally for many genes, and elongation rate influences splice site recognition, spliceosome assembly, and alternative splicing[83]. Alternative splicing allows for several mRNA products to be produced from the same gene. Binding of splicing enhancers by serine/arginine rich (SR) proteins or silencers by heterogeneous nuclear RNPs can influence exon inclusion[81,84]. This regulation contributes to the tissue- or cellular response-specific expression of certain RNA isoforms[85]. Alternative splicing events can also lead to the introduction of a premature stop codon, which will result in targeting of the mRNA for degradation through the nonsense-mediated decay (NMD) pathway[86]. Therefore, regulation of alternative splicing can be used as a mechanism to manage transcript abundance.

Most mRNAs are polyadenylated at the 3'end of the transcript. After cleavage of the nascent RNA downstream of the PAS, a non-templated poly(A) tail is added by a poly(A) polymerase[81,87]. The poly(A) tail aids in protection of the 3'end and RNA export into the cytoplasm for translation. Alternative polyadenylation occurs when different cleavage sites produce mRNAs with different 3'UTRs[84]. This can result in changes in mRNA stability, localization, or transport. Changes in the 3'UTR of the transcript can affect post-transcriptional regulation of the mRNA, such as targeting by miRNAs[88]. Therefore, regulation of polyadenylation allows for cell type-specific control over gene expression.

After processing, mature mRNAs assemble with export factors into complex messenger ribonucleoprotein (mRNP) particles[89,90]. Export receptors are recruited and transport mRNPs through the nuclear pore complex. Once in the cytoplasm, mRNA is released for translation. RNA stability in the cytoplasm also influences transcript abundance and gene expression levels. Degradation rates of mRNA can be specific to a cell type or cellular response. RNA binding proteins and non-coding RNAs can bind to regulatory elements on the RNA to affect stability by targeting it for or shielding it from degradation[91]. AU-rich elements (AREs) are well-characterized regulatory regions in the 3'UTR of transcripts[92]. Proteins which bind to AREs can influence mRNA stability and translation through altering local RNA structure[93,94] or recruiting degradation machinery[91,92]. Noncoding RNAs (ncRNAs) also influence mRNA stability by blocking translation or targeting mRNAs for degradation[95].

A major mechanism for mRNA decay is the deadenylation-dependent pathway[91,96]. The poly(A) tail is removed by a deadenylase. Then, degradation of the transcript can occur in the 3'-5' direction via the exosome. When a few 5' nucleotides remain, the cap is removed by the scavenger decapping enzyme. Alternatively, after deadenylation, decapping can occur to allow degradation in the 5'-3' direction by an exoribonuclease. A second pathway for RNA decay is the endoribonucleolytic pathway, which is deadenylation-independent[91,97]. Endoribonucleases often target transcripts undergoing active translation, and cleave the mRNA in the 3'UTR. The RNA is then subject to 3'-5' degradation or decapping and 5'-3' degradation. Processing bodies (P-bodies) are cytoplasmic foci which contain mRNA degradation proteins[98]. In addition to being sites of mRNA decay, transcripts may also exit P-bodies for translation, and so they may also play a role in mRNA sorting and storage.

1.2 Bru-seq techniques

The study of gene expression and the detection of RNA go hand in hand. The development of the Northern blot method in 1977 enabled scientists to detect specific

RNA species within a sample using probe hybridization[99]. Reverse transcriptase PCR then allowed detection of low-abundance RNAs through cDNA production and amplification[100]. One important limitation of these methods is that they only allow researchers to assess a handful of selected genes at a time. Development of the microarray enabled researchers to simultaneously evaluate gene expression of thousands of genes at once. DNA microarrays utilize probes on a chip to determine relative abundance of sequences in a sample and have been widely used for transcriptional profiling[101]. However, the microarray approach still relies on prior knowledge of target sequences for probe design and does not provide additional sequence information for transcripts.

The emergence of high throughput sequencing technology allowed for analysis of genome-wide transcription. RNA-seq, which involves massively parallel cDNA library sequencing, allows for a more unbiased analysis of RNA detection compared to microarray analysis[102]. RNA-seq does not require genomic sequence information to detect transcripts and can reveal transcript sequence variations[103]. Expression levels are determined based on the number of reads detected for a transcript. Because either total RNA or poly(A)+ RNA is used for generating RNA-seq libraries, this technique commonly provides an assessment of steady state RNA levels within cells[103]. The steady state population is the product of both RNA synthesis and stability, however RNA-seq does not distinguish how each of these events is contributing to transcript abundance. A clearer understanding of how RNA synthesis contributes to transcript homeostasis requires an evaluation of nascent transcription in the cell. With this goal in mind, the Ljungman lab developed the Bru-seq technique, as well as complementary modifications, to study nascent RNA expression in live cells.

1.2.1 Bru-seq

In order to capture transcripts which are being actively transcribed in cells, our lab metabolically labels RNA. This involves the use of a uridine analog, bromouridine (Bru), which is incorporated into newly synthesized RNA by RNAPII when it is added to cell media. After RNA extraction, Bru-labeled RNA can be isolated from total RNA using anti-BrdU antibodies conjugated to magnetic beads. The Bru-labeled RNA is used to produce strand-specific cDNA libraries which are then sequenced. Transcriptional analysis is performed using mapped read density to the reference genome. This technique is called Bru-seq[104,105]. Bru-seq specifically provides a genome-wide picture of RNA synthesis by distinguishing nascent RNA from previously synthesized RNA.

In typical Bru-seq experiments, cells are pulse-labeled with Bru for 30 minutes and then immediately lysed in TRIzol reagent to preserve the RNA. For transcribed genes, we obtain mapped reads across the entire gene. This is in contrast to RNA-seq samples which contain mostly mRNA and have an enrichment of reads within exonic sequences. Based on read counts within a gene, and normalization for gene length and total library read count, we can calculate expression level of a gene (in reads per thousand basepairs per million reads, RPKM). In addition to detecting expression levels of annotated genes and ncRNAs, Bru-seq can also identify unannotated transcripts such as lncRNAs. Comparison of Bru-seq data across several cell lines revealed that many of these unannotated lncRNAs display cell type-specific expression patterns[105].

The initial Bru-seq experiment was performed on normal human fibroblasts, and estimated that approximately 34% of the genome produced nascent transcripts in these cells[104]. Most of the sequence reads mapped to introns, while the majority of the remaining reads mapped to either exons or unannotated intergenic regions. By comparing the transcriptional profiles of normal and treated or mutant cells, Bru-seq can identify transcriptional changes related to a cellular response or certain mutation. To explore changes in transcription following cellular stimulation, normal fibroblasts were treated with tumor necrosis factor (TNF) to induce the inflammatory response[106]. Bru-seq detected 472 genes that were upregulated and 204 genes that were downregulated at least 2-fold following incubation with TNF for 60 minutes. Functional annotation analysis of upregulated genes indicated enrichment of genes involved in inflammatory response pathways. Overall, Bru-seq provides a way to analyze RNA synthesis by capturing nascent transcripts and allows for assessment of transcriptional changes at the level of RNA production.

1.2.2 BruChase-seq

Because Bru-seq utilizes metabolic labeling, the Bru labeling period can be followed by a chase period with uridine. By applying chase periods of different lengths of time, we can analyze RNA populations of distinct ages. This modification of the Bru-seq technique is called BruChase-seq[104,105]. While Bru-seq detects levels of RNA synthesis, BruChase-seq indicates how relative levels change after a given period of time.

In typical BruChase-seq experiments, cells are Bru-labeled for 30 minutes followed by a 6 hour uridine chase period. For transcribed genes, we observe an enrichment of mapped reads within exons. This is consistent with expected maturation and splicing of RNA occurring during the chase period. Relative stability of a transcript can be calculated by comparing the RPKM value from exons in the BruChase-seq data to the RPKM value from the entire gene in the Bru-seq data. Demonstrating the power of BruChase-seq, transcripts from genes bearing nonsense or frameshift mutations displayed lower stability compared to transcripts from wildtype versions of the same genes[105]. These results are likely due to increased RNA degradation through the nonsense-mediated decay pathway. Conversely, increased stability of the *MYC* oncogene was observed in certain cancer cell lines, which may contribute to MYC overexpression in human tumors. BruChase-seq is useful for assessing how RNA stability contributes to transcript abundance.

In the initial BruChase-seq experiment on normal human fibroblasts, there did not appear to be a relationship between relative expression levels and relative RNA stability[104]. This supports the notion that RNA synthesis and degradation independently influence transcript abundance. Functional annotation analysis revealed that while ribosomal genes were highly transcribed, they were also highly unstable. BruChase-seq was also performed following TNF treatment to examine changes in transcript stability[107]. TNF treatment resulted in at least a 2-fold increase in transcript stability of 152 genes and at least a 2-fold decrease in stability of 58 genes. Genes involved in the inflammatory response demonstrated dramatically increased transcript stability, and many of these also had upregulated synthesis. These results may represent TNF activation of both gene induction and RNA stabilization, or may be a result of reduced

RNA decay due to increased transcript levels overwhelming the degradation machinery. Overall, Bru-seq and BruChase-seq data revealed complex regulation of TNF response genes at either or both the transcriptional and post-transcriptional level.

In addition to the assessment of stability, BruChase-seq can also be used to examine splicing dynamics. By analyzing different ages of RNA, we can follow the progression of splicing over time. As the time of the chase period increases, the amount of signal within intronic sequence decreases. Initial observations indicate that the rate of splicing for introns both within and across genes is not uniform. Additionally, certain transcripts appear to contain retained introns even after the 6 hour chase period[104]. Reads within these retained introns mapped across intron-exon boundaries. This indicates that detection of these introns represents inclusion in the transcript rather than decreased degradation of spliced introns. Overall, BruChase-seq is a powerful tool for studying posttranscriptional RNA processing and decay separately from RNA synthesis.

1.2.3 BruDRB-seq

The BruDRB-seq technique was developed to study genome-wide elongation rates[68]. The arrest of RNAPII at promoter-proximal sites can be achieved through treatment of cells with 5,6-dichlorobenzimidazole 1-β-D-ribofuranoside (DRB) [108]. This drug inhibits the transition of RNAPII into productive elongation but does not hinder transcription initiation or polymerases that are already elongating across the gene body. These effects are reversible and removal of DRB from cell media results in the release of stalled polymerases into productive elongation. Thus, DRB treatment allows us to synchronize transcription of genes at the TSS. For the BruDRB-seq protocol, we pre-treated cells with DRB for 60 minutes. Bru-labeling during the last 10 minutes of treatment revealed severely reduced read levels across genes, reflective of transcription elongation inhibition. Removal of the drug followed by Bru-labeling for 10 minutes resulted in a peak of transcriptional reads at the TSS of genes. In cells that were labeled for 20 minutes, we observed movement of the synchronized transcriptional wave further along the gene. Because DRB does not inhibit polymerases which are already in the process of productive elongation, treatment results in a clearing of polymerases from the gene that is dependent on gene length and time. For large genes over 200kb, we observed a retreating wave of transcription at the end of genes following DRB treatment. This wave also moved towards the end of the gene during later periods following treatment.

To determine elongation rate within a given gene, we sought to measure the distance travelled by RNAPII during the labeling period based on the leading edge of the transcriptional wave. To do this in a genome-wide manner, we developed an inference model based on a three-state hidden Markov model (HMM). The purpose of the HMM was to distinguish the region immediately upstream of the TSS, the region of the advancing transcriptional wave, and the region of reduced transcription downstream of the wave. Using this method, an elongation rate was calculated for all expressed genes over 40kb.

We used BruDRB-seq to analyze genome-wide elongation rates in five cell lines. Our data revealed that genes exhibited a broad range of elongation rates, with median values for the cell lines ranging from 1.25 to 1.75kb/min. Comparisons of elongation rates for genes across cell lines showed that certain genes elongate at similar relative rates while other genes showed variable elongation rates in the different cell lines. We identified several gene features which were associated with elongation rate, including longer gene length, low DNA complexity, and greater distance from other expressed transcripts. High exon density and GC content correlated with slower elongation rates. Because we identified genes which appeared to have cell type-specific elongation rates, we also explored whether any chromatin modifications were associated with elongation rate. We compared our BruDRB-seq data to available ENCODE ChIP-seq data and found that the density of the histone marks H3K79me2 and H4K20me1 correlated with elongation rate. We did not find a correlation between elongation rate and H3K36me3 or RNAPII density, which are known to correlate with gene expression levels. This suggested that high elongation rates were not simply a reflection of high expression levels. Taken together, BruDRB-seq is an effective method for measuring genome-wide elongation rates, and revealed that both genomic features and epigenetic marks influence elongation rate. This data reveals that chromatin modifications are a potential mechanism through which cells can fine-tune elongation rates in order to influence expression timing. This regulation may be important for coordinating transcriptional timing at times of active replication during the cell cycle or in response to environmental stimuli.

1.2.4 BruUV-seq

The final Bru-seq modification which was developed is called BruUV-seq[109]. Treatment of cells with UVC light results in the formation of pyrimidine dimers and 6-4 photoproducts within DNA[110]. These DNA lesions block transcriptional elongation by causing RNAPII to stall[111,112]. In BruUV-seq, cells are UV irradiated prior to Bru-labeling. This results in a redistribution of reads within genes, with a large peak near the TSSs and fewer reads across the gene body. This effect is dose-dependent, with higher doses of UV leading to higher and narrower peaks at TSSs. Therefore, BruUV-seq allows for the identification of TSS usage within cells.

For the initial BruUV-seq experiment, we compared Bru-seq and BruUV-seq data with histone modification ChIP-seq data from ENCODE in similar cell lines[113]. For individual genes, we found that the BruUV-seq peaks at TSSs corresponded to H3K4me3 peaks, a mark associated with promoters[114]. Genome-wide analysis revealed that while approximately 95% of expressed genes displayed a single TSS BruUV-seq peak, there were genes with up to 5 distinct TSS peaks. Furthermore, in a comparison of two different cell lines, we observed that TSS usage was cell type-specific for certain genes. While H3K4me3 peaks were seen at all putative promoters for a gene, BruUV-seq allowed for identification of TSSs which were being actively used for transcription initiation in a given cell line. Additionally, it appeared that certain clusters of miRNA and small nucleolar RNA genes were transcribed from a single TSS, suggesting polycistronic transcription of these genes. We went on to use BruUV-seq to examine changes in gene expression following TNF induction of the inflammatory response. Changes in BruUV-seq signal intensity at the TSS corresponded with changes in expression levels in the Bru-seq data, demonstrating the usefulness of this technique to distinguish how multiple TSSs may be individually contributing to RNA synthesis. Because UV treatment also resulted in enrichment of normally unstable RNAs such as PROMPTs and eRNAs, BruUV-seq can be used to annotate putative enhancer elements, which is the focus of Chapter IV. Overall, BruUV-seq allows for the identification of active TSSs and enhancers within a cell, and can be used to analyze changes in TSS and enhancer usage during a cellular response.

1.3 Other nascent RNA sequencing technologies

A variety of other techniques have been developed during the past five years to analyze the nascent transcriptome in order to study transcriptional processes. This section describes these techniques and discusses potential advantages and disadvantages.

One method to evaluate nascent RNA populations is to only sequence RNA found in certain cellular fractions. Biochemical fractionation of cells can be used to isolate RNA bound to chromatin or within transcription factories, which will be enriched for nascent transcripts compared to total RNA[115,116]. However, analysis of chromatin-bound RNA fractions have suggested that completed, cleaved transcripts for many genes accumulate on chromatin[117]. Additionally, ncRNAs such as lncRNAs and snoRNAs are found to be in chromatin and transcription factory fractions due to their roles in transcriptional regulation or RNA processing[116,118]. Therefore, while these methods can be used to enrich for nascent RNA, they will still capture RNA transcripts that are associated with these cellular components regardless of the age of the RNA.

Another method to look at RNA which is being actively transcribed is to capture transcriptionally engaged RNAPII through immunoprecipitation, and then sequence the 3'end of the associated RNA, or NET-seq[119]. Because the RNA-DNA-RNAPII complex is extremely stable, these complexes can be purified directly from live cells without crosslinking. One advantage of NET-seq is that it can map the position of RNAPII at nucleotide resolution[119,120]. All transcriptionally engaged RNAPII will be detected and mapped, including paused polymerases. NET-seq is also able to capture unstable transcripts such as PROMPTs. However, because NET-seq requires immunoprecipitation of the RNAPII complex, this technique is relatively challenging, potentially requires a large number of cells, and cannot be used to age and follow RNA over time.

A currently popular method for nascent RNA sequencing is GRO-seq, which uses nuclear run-on assays to extend nascent RNAs associated with transcriptionally engaged polymerases[121]. In this method, nuclei are extracted and RNAPII is allowed to run on in the presence Br-UTP. These *in vitro* transcribed RNAs are then isolated using an antibody and used for sequencing. The original GRO-seq method was used to study promoter-proximal pausing and used the detergent sarkosyl to release paused RNAPII during the run-on assay[121]. A variation of the method without sarkosyl treatment can be used to assess actively elongating RNAPII[122]. Another modification of this method that restricts the number of incorporated bases, PRO-seq, can be used to map transcriptionally-engaged RNAPII at nucleotide resolutions[123]. GRO-cap, which enriches for capped RNAs, can be used to identify TSSs and also captures unstable transcripts such as PROMPTs and eRNAs[124,125]. GRO-seq and its related variations allow for high resolution and efficient detection of transcriptional activity. Nevertheless, this technique requires large amounts of material and is a relatively difficult technique. Furthermore, because this method involves *in vitro* transcription, this might introduce artifacts that may not be observed during *in vivo* transcription.

Similar to the Bru-seq techniques, other forms of metabolic labeling to isolate nascent RNA have been applied. Other analogs that have been used for labeling include 4-thiouridine (4sU) and 5-ethynyluridine (EU). These analogs are efficiently incorporated into nascent RNA, however in contrast to Bru-labeling, prolonged exposure to both these analogs has been shown to result in reduced cell viability[126]. Labeling with 4sU along with whole genome sequencing has been effectively used to study genome-wide RNA synthesis[127], RNA stability[126,127], splicing kinetics[128], and transcription elongation[70].

1.4 Dissertation objectives

Each step of the transcriptional process is highly regulated and contributes to the overall control of gene expression. Transcription is often studied using methods that assess RNA abundance of full-length, processed transcripts. These methods limit studies of transcriptional timing to transcript completion and cannot precisely assess when transcript initiation is induced or repressed. Additionally, it can be difficult to distinguish how initiation and elongation are individually contributing to genome-wide expression patterns. Furthermore, unstable transcripts such as those produced at enhancer elements may be missed, making it challenging to simultaneously examine enhancer activity and gene expression.

The goal of my thesis work was to further understand the role of early transcriptional events in transcriptional timing and the regulation of gene expression. Through the use of the Bru-seq nascent RNA sequencing technology, I was able to study global transcriptional dynamics at the levels of productive initiation, elongation, and enhancer activation. This dissertation is organized according to the following specific aims:

Specific Aim 1 (Chapter II): Examine genome-wide early transcriptional dynamics during the serum response. Bru-seq can capture immediate induction or repression of genes following serum stimulation, revealing the temporal dynamics of the transcriptional response to serum. This novel dataset provides genome-wide information about transcriptional timing at the level of productive initiation and reveals the complexity of transcriptional patterns occurring during this cellular response.

Specific Aim 2 (Chapter III): Explore the impact of gene length on transcriptional timing and its role in establishing temporal gene expression patterns during the serum response. RNA synthesis across long genes can be monitored using Bru-seq in order to follow the progression of transcription after serum addition, and to study the effects of gene length on transcriptional timing of response genes. While it is currently unclear why extremely long genes have been maintained during evolution, this analysis provides evidence for a role of gene length in establishing proper temporal gene expression patterns during the serum response.

Specific Aim 3 (Chapter IV): Identify active enhancer elements based on eRNA production and investigate the impact of inhibition of the enhancer chromatin regulator BRD4 on enhancer activity. BruUV-seq can detect enhancer RNA and indentify changes in the production of this RNA following drug treatment. This dataset demonstrates the importance of studying the effects of drugs that target global chromatin regulators such as BRD4, and provides important information for studying enhancer biology.

This work uses the nascent sequencing technology Bru-seq to overcome limitations of other techniques that simply assess RNA abundance and provides greater insight into the mechanisms of transcriptional regulation. The datasets from these experiments provide a valuable resource for the research community to pursue future studies of the regulatory mechanisms contributing to gene expression.

CHAPTER II

Capturing the dynamic transcriptome during the serum response[1]

2.1 Abstract

Dynamic regulation of gene expression is of fundamental importance during many biological processes such as cell state transitioning, cell cycle progression, and stress responses. Activation of specific transcription factors or repressors through signal transduction pathways leads to rapid initiation or repression of target gene transcription. In this study, we used serum stimulation as a cell response paradigm to apply the nascent RNA Bru-seq technique in order to capture early changes in the nascent transcriptome. Our data provides an unprecedented view of the dynamics of genome-wide transcription during the first two hours of serum stimulation in human fibroblasts. While some genes showed sustained induction or repression, other genes showed transient or delayed responses. As expected, early response genes such as those encoding components of the AP-1 transcription factor were immediately but transiently induced. Surprisingly, transcription of important DNA damage response genes was rapidly repressed. These results provide a unique genome-wide depiction of the dynamic induction and repression patterns of serum response genes, and they demonstrate the utility of Bru-seq to comprehensively capture rapid changes of the nascent transcriptome.

2.2 Introduction

Signal transduction cascades are critical regulators of cell processes such as differentiation, proliferation, and stress responses. These cascades activate preexisting transcription factors, which in turn induce various groups of target genes. Some of these target genes are also transcription factors that, after the completion of transcription and

[1] The data presented in Chapter II have been submitted as a manuscript and accepted by Biology Open: Kirkconnell, K.S., Paulsen, M.T., Magnuson, B., Bedi, K., Ljungman, M. Capturing the Dynamic Nascent Transcriptome during Acute Cellular Responses: The Serum Response. *Biology Open*. (2016). In Press.

translation, go on to regulate their own sets of target genes. These series of regulatory and transcriptional networks allow for an initial triggering event to produce complex global gene expression changes in a temporal and dynamic way.

Growth factor activation is a widely used system for studying rapid changes in gene expression. Fibroblasts grown in serum-free media enter a G_0/G_1 state, and subsequent addition of serum to these cells results in global gene expression changes as cells prepare to progress through the cell cycle again[129]. Studies analyzing the genome-wide early transcriptional response to serum stimulation indicate that response genes exhibit various temporal expression patterns[130-132]. In addition to recruitment and initiation of transcription by RNA polymerase II (RNAPII), several additional mechanisms are thought to influence gene expression timing, including release of paused polymerases, elongation progression, termination, and polyadenylation[56]. These various points of regulatory control, which differ for individual response genes, make it challenging to unravel the complex global patterns of transcriptional activation and repression.

Methods such as microarrays and RNA-seq are excellent for detecting polyadenylated mRNA or changes in the total mRNA pool, however the detection of these changes may not directly reflect the initial transcriptional response. Therefore, we used Bru-seq, a nascent RNA sequencing technique, to focus on transcriptional changes near transcription start sites (TSSs)[104,105]. By measuring changes in the nascent RNA production emanating from TSSs, we could identify genome-wide dynamic changes in productive transcription initiation. Our data indicate that there are several distinct gene expression patterns that occur in response to serum stimulation, and genes in distinct cellular pathways often exhibit similar response patterns. This study was able to capture the global dynamics of RNA production in unparalleled detail during the early serum response and showcases the utility of Bru-seq in monitoring genome-wide transcriptional dynamics.

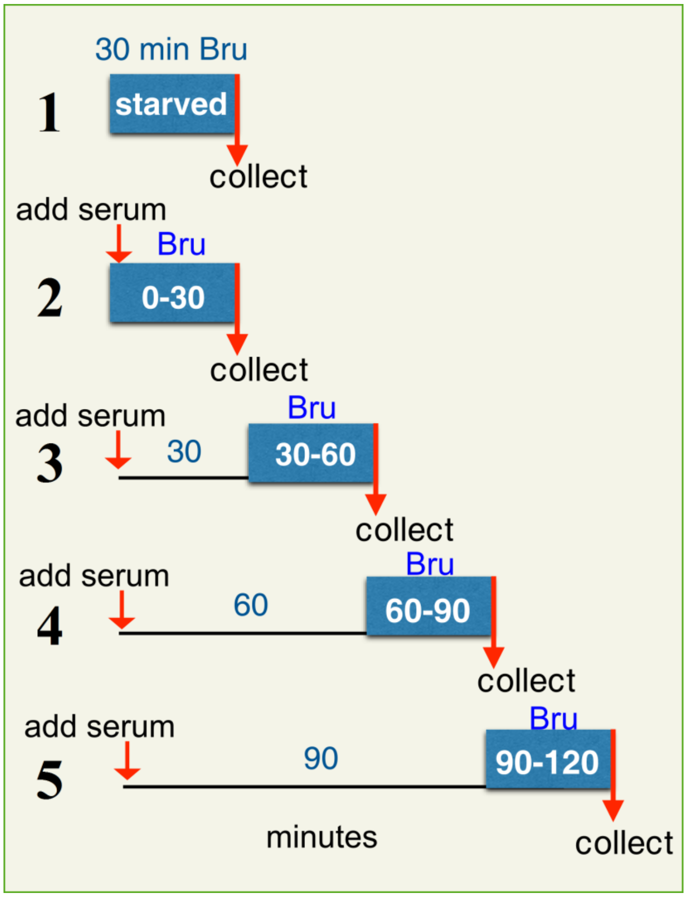


Figure 2. 1: Bru-seq experimental outline.

Cells were maintained in serum-free media for 48 h prior to labeling. Bromouridine labeling of nascent RNA was performed for 30 minute periods (shown in blue) on starved human fibroblasts (1), or at different times after serum addition (2-5).

2.3 Results

2.3.1 Immediate, sustained effects of serum stimulation on transcription initiation

For this study, human fibroblasts were serum starved for 48 hours prior to serum addition in order to activate the serum response. We added bromouridine (Bru) to cell media for 30 minutes to label nascent RNA in starved and serum stimulated cells (Fig. 2.1). For serum stimulated cells, we Bru-labeled cells at various timepoints following serum addition. Using Bru-seq to monitor nascent RNA production allowed us to capture the dynamic landscape of transcriptional alterations during the first 2 hours of the serum response. By mapping the reads for each 30-minute labeling period, we were able to visualize active transcription across individual genes. To focus on changes in productive initiation, we measured the abundance of reads within the whole gene for those 30 kb or shorter, and within the first 30 kb for longer genes. We chose a 30 kb-window because we expected that based on an average elongation rate of 1.4 kb/min and the 30 minute labeling period, RNAPII would have fully traversed the sampling area[68]. Importantly, this approach allowed us to more accurately assess the rate of transcription for very large genes, which may have been severely underestimated if transcription within the full length of the gene was considered.

Addition of serum, which contains growth factors, induces the expression of many different types of genes including those that encode transcription factors, proteins involved in cell movement such as cytoskeletal and extracellular matrix proteins, and signaling factors such as cytokines[133]. Indeed, using Bru-seq we observed the upregulation of genes with related functions following serum addition. Certain genes, such as the cytoskeleton gene *TPM1*, were immediately upregulated following serum addition, which was visible through an increase in transcriptional reads across the entire gene (Fig. 2.2A). This increase was observed during each labeling period. The cell adhesion gene *FERMT2* behaved in a very similar manner, but because this gene is longer (~90kb versus ~30kb) we only observed reads at the 5'-end of the gene during the first labeling period (Fig. 2.2B). However, during the subsequent labeling period (30-60min), the wave of nascent transcription had reached the 3'-end of the gene. One

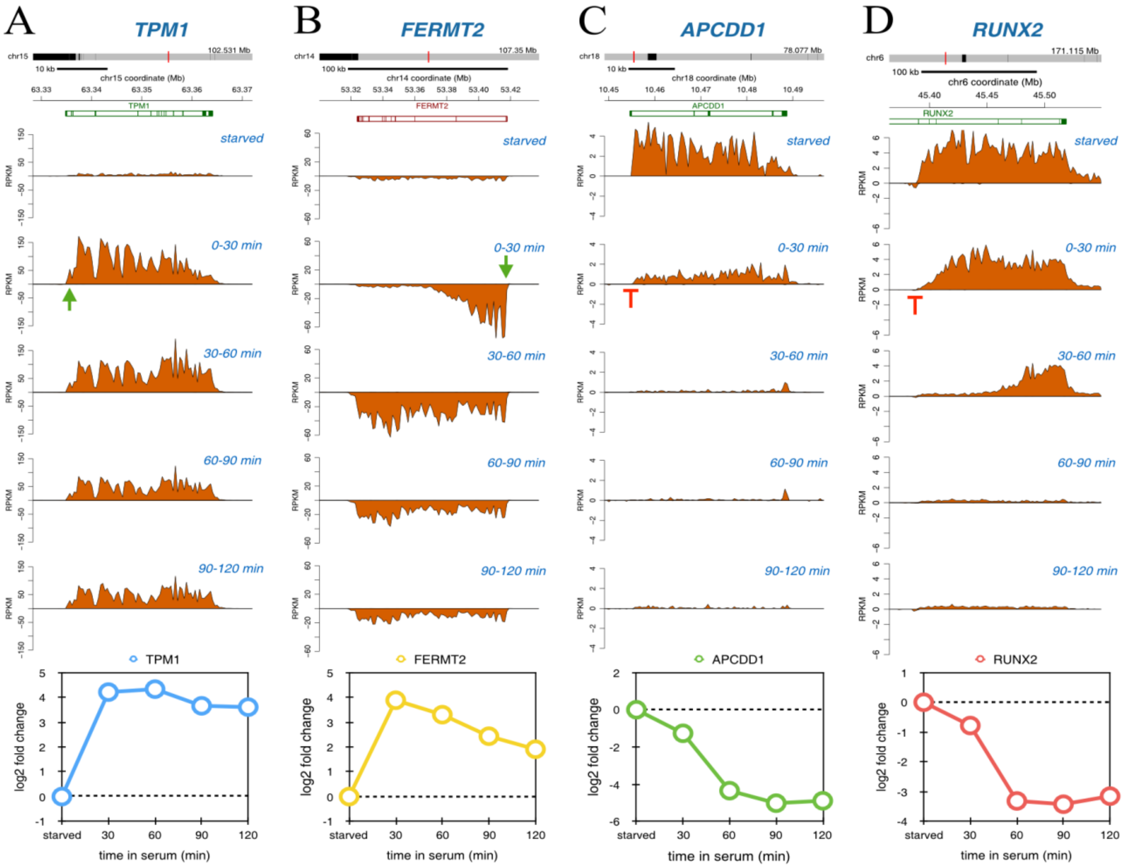


Figure 2. 2: Immediate serum-response genes.

Bru-seq traces for *TPM1* (A), *FERMT2* (B), *APCDD1* (C) and *RUNX2* (D) during starved conditions and different periods after serum stimulation. Genes are shown at the top in in green and red, with exons indicated by vertical bars. Transcription induction is indicated by a green arrow and transcription repression is indicated by a red T. The positive y-axis represents plus-strand signal and the negative y-axis represents minus-strand signal, hence plus-strand genes transcribe from left to right and minus-strand genes transcribe from right to left. The graphs at the bottom depict the log$_2$ fold change values calculated within the first 30 kb of the genes for each labeling period.

advantage of Bru-seq over traditional RNA-seq, which measures steady-state RNA, is that it can capture rapid repression of transcription very efficiently since it does not rely on the degradation of preexisting RNA for the detection of inhibition. For example, the signaling gene *APCDD1* was actively transcribed during serum starvation but was repressed upon serum addition (Fig. 2.2C). This decrease in reads was maintained throughout the following labeling periods. The *RUNX2* transcription factor gene was also repressed in response to serum, and due to its considerable length (>100 kb), transcription at the end of the gene was unaffected until 60-90 minutes after serum addition (Fig. 2.2D). Thus, even though genes can be induced or repressed immediately following serum stimulation, there is a time delay before the generation or loss of full-length products that is proportional to gene length.

2.3.2 Transient effects of serum stimulation on transcription initiation

While the previously described genes displayed sustained transcriptional induction or repression following serum stimulation, other genes demonstrated transient regulation. For example, the transcriptional activator gene *NR4A3* was immediately induced after serum addition, but began to decrease transcription during the 60-90 minute labeling period (Fig. 2.3A). The translational repressor gene *SAMD4A* was also transiently induced by serum, and initiation levels returned to serum-starved levels after 60-90 minutes. Because this gene is very long (>200 kb), there was a delay before the 3' end of the gene experienced the effects of this brief pulse of transcriptional induction (Fig. 2.3B). We also observed genes such as the cell junction gene *KIRREL* (Fig. 2.3C) and the metabolism gene *ABCA1* (Fig. 2.3D) which showed brief inhibition of transcription followed by a return to baseline expression after serum addition. Again, effects were delayed in distal gene regions according to length.

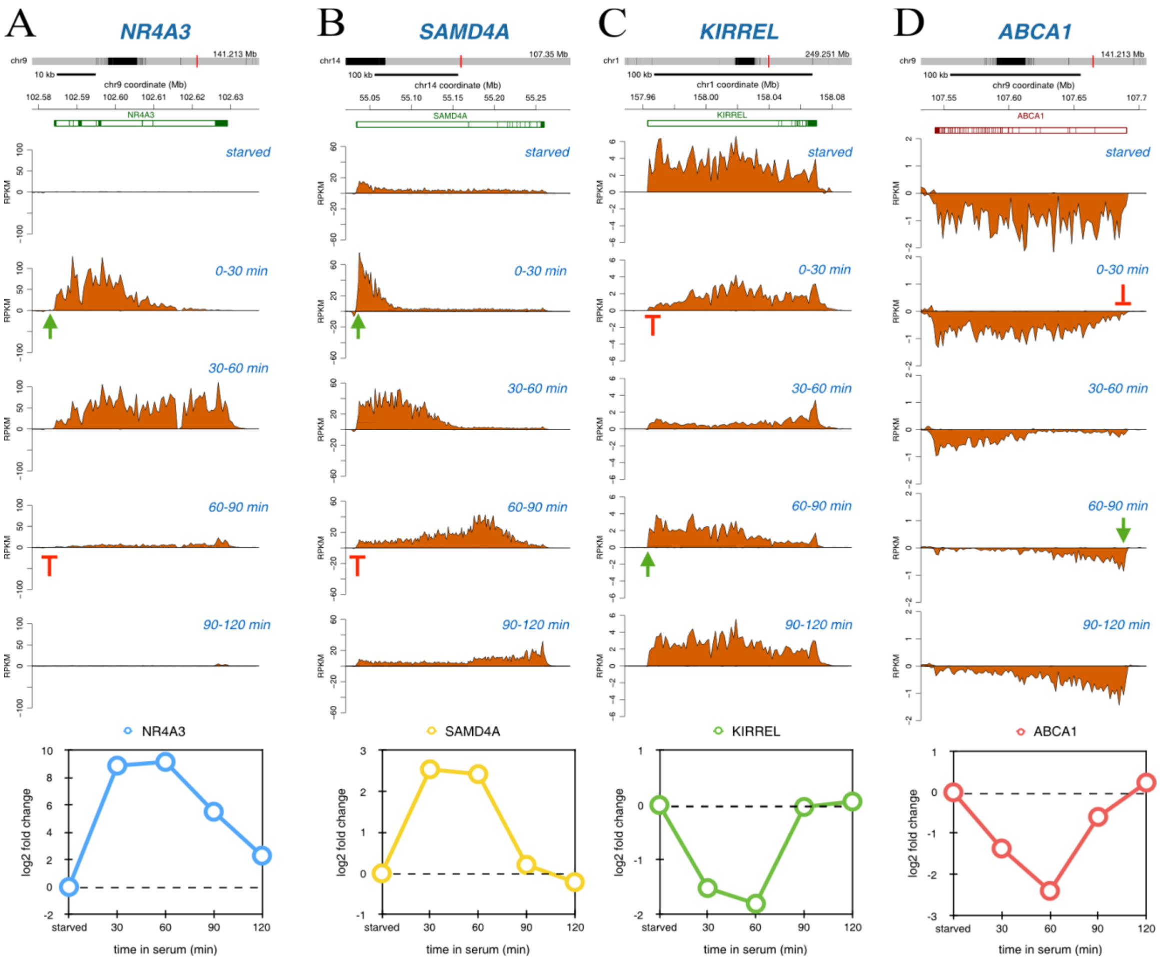


Figure 2. 3: Transient serum-response genes.

Bru-seq traces for *NR4A3* (A), *SAMD4* (B), *KIRREL* (C), and *RUNX2* (D) during starved conditions and different periods following serum addition. Data representation as in Fig. 2.2.

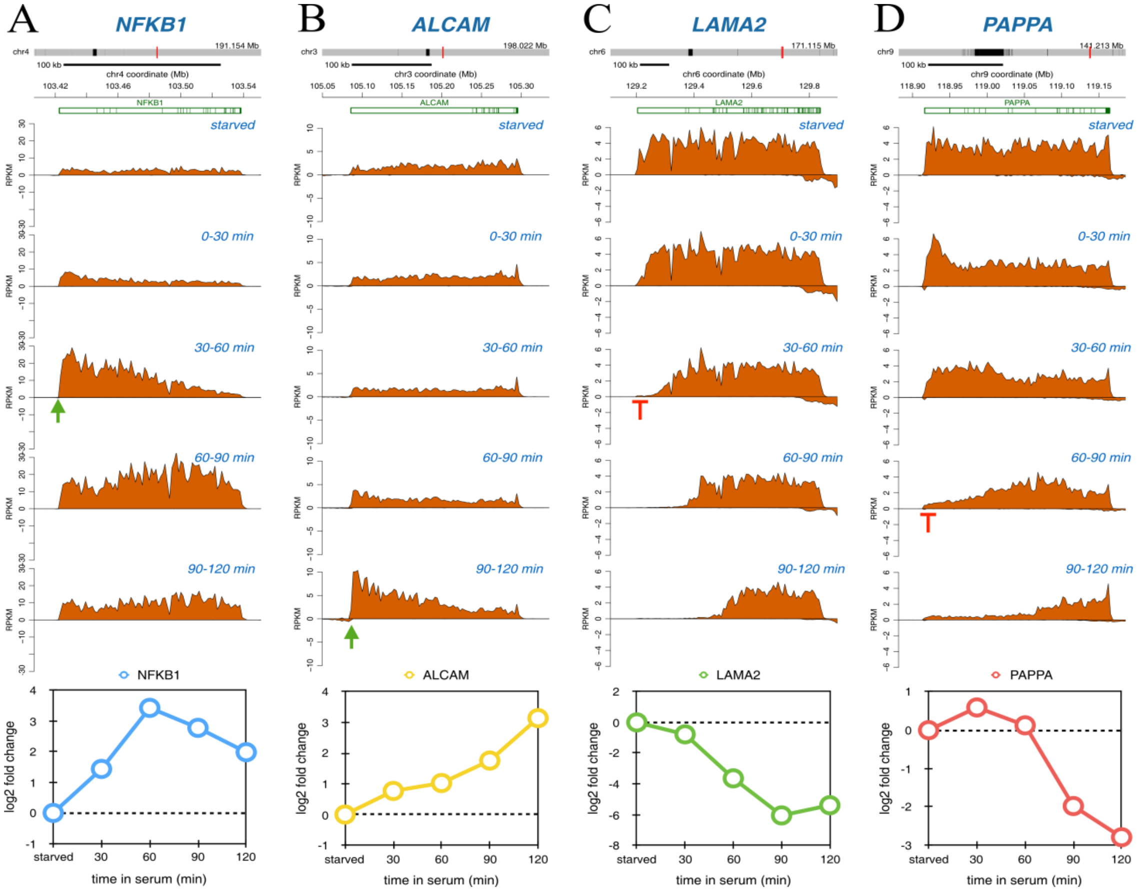


Figure 2. 4: Delayed serum-response genes.

Nascent RNA sequencing reads for *NFKB1* (A), *ALCAM* (B), *LAMA2* (C), and *PAPPA* (D) during starved conditions and different periods following serum addition. Data representation as in Fig. 2.2

2.3.3 Delayed effects of serum stimulation on transcription initiation

Certain gene responses occurred during later labeling periods following serum activation. Transcriptional induction of the transcription factor gene *NFKB1* peaked around the 30-60 minute labeling period (Fig. 2.4A) and induction of the cell migration gene *ALCAM* peaked during the 90-120 minute labeling period (Fig. 2.4B). Considerable repression of the extracellular matrix gene *LAMA2* began 30-60 minutes after serum addition (Fig. 2.4C), and the wound healing gene *PAPPA* exhibited repression 60-90 minutes after serum addition (Fig. 2.4D). These delayed responses may be regulated by transcription activators or repressors that are transcriptionally induced as part of the immediate response to serum.

2.3.4 Genome-wide patterns of transcriptional regulation following serum stimulation

After observing this variety of transcriptional responses, we went on to perform a genome-wide analysis of serum-induced transcriptional changes. Using a 2-fold change cutoff, we observed 1417 genes that were upregulated (Fig. 2.5A) and 636 genes that were downregulated (Fig. 2.5F) during at least one labeling period following serum stimulation. We classified serum response genes based on transcription patterns near the TSS during the first 2 hours following serum stimulation using the following categories: sustained induction or repression (Fig. 2.5B,G), induction or repression followed by a return to baseline expression (Fig. 2.5C,H), delayed induction or repression (Fig 2.5D,I), and both induction and repression (Fig. 2.5E,J). These results illustrate the intricate patterns in which cells regulate transcription following a global cellular response, and demonstrate the power of Bru-seq in capturing rapid and dynamic changes in the nascent transcriptome.

2.3.5 Functional analysis of serum response genes

We utilized the DAVID functional annotation tool to explore whether the serum response genes identified by Bru-seq were enriched for certain functional pathways, including those known to be related to serum activation[134,135]. First, we focused on our group of transcriptionally induced genes to identify potential pathways which were upregulated in response to serum activation. We found significant enrichment of genes

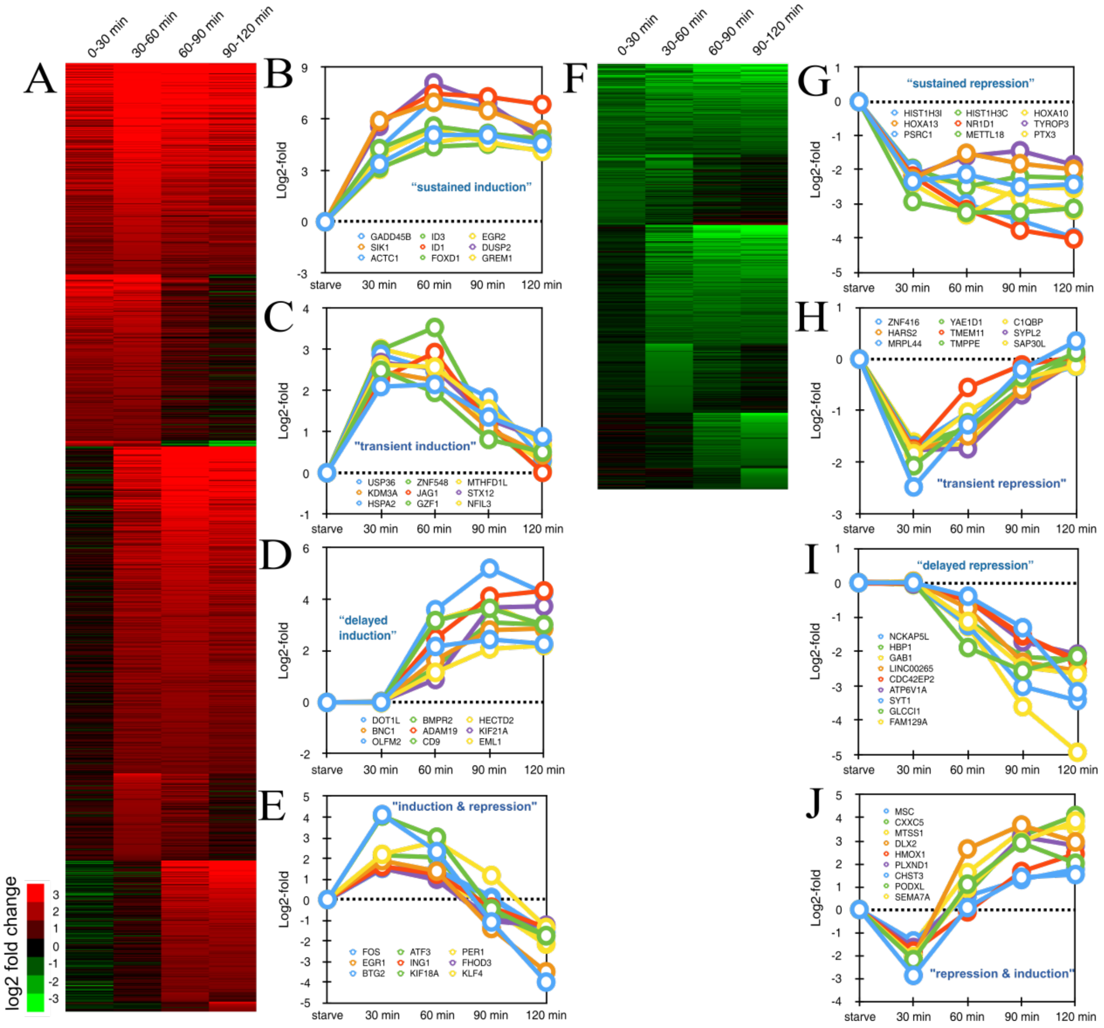


Figure 2. 5: Global dynamics of the nascent transcriptome following serum stimulation.

Heatmaps of induced (A) and repressed (F) genes in response to serum stimulation based on log2-fold change values within the first 30 kb of genes. Examples of gene groupings exhibiting various transcriptional patterns are shown in graphs (B-E) and (G-J).

related to cellular structures such as “extracellular matrix”, “focal adhesions”, “actin cytoskeleton”, and “tight junctions” (Fig. 2.6A), which are known to be important during the serum and fibroblast wound response[130,133,136]. Genes involved in various signaling pathways such as toll-like receptor, chemokine, TGFβ, and MAPK signaling were also enriched following serum stimulation. Enrichment of genes implicated in these pathways was seen during each labeling period.

Next, we examined our set of transcriptionally repressed genes to identify potential pathways which were downregulated in response to serum activation. We observed enrichment of genes involved in signaling pathways including MAPK, ErbB, Jak-STAT, and cytokine receptor signaling (Fig. 2.6B), however the enrichment of these genes was less significant than the upregulated signaling pathway genes. While the genes involved in these pathways were enriched during each labeling period, genes in other pathways had higher enrichment scores only during the later labeling periods. These pathways included “nucleotide excision repair”, “cell cycle”, “WNT signaling”, “aminoacyl tRNA biosynthesis”, and “p53 signaling” (Fig. 2.6C).

Additionally, we performed a gene set enrichment analysis (GSEA) for each labeling period[137]. Expressed genes during a given period were ranked according to fold change in expression compared to starved cells, and this ranked gene list was used to calculate normalized enrichment scores (NESs) for gene sets. Gene sets with high NESs during each labeling period, indicating sustained increased transcription levels, included genes induced by NRG1, EGF, TGFβ, TNF, WT1, KRAS, and the inflammatory response (Fig. 2.6D). Gene sets with low NESs during each labeling period, indicating sustained decreased transcription levels, included those related to “meiosis”, “telomere and chromosome maintenance”, “RNA polymerase I transcription”, and “amyloids” (Fig. 2.6E).

2.3.6 Rapid induction ofAP-1 transcription factor genes following serum stimulation

The AP-1 transcription factor is primarily comprised of Jun, Fos, and ATF protein dimers[138]. *FOS* and *JUN* are known to be rapidly and transiently induced following serum stimulation[133]. Bru-seq revealed that several AP-1 related genes were rapidly but transiently induced following serum addition (Fig. 2.7A-G). The mechanism behind this

A **induced**

pathway	0-30min ES	0-30min p-value	30-60min ES	30-60min p-value	60-90min ES	60-90min p-value	90-120min ES	90-120min p-value
ECM receptor interaction	27.34	4.38E-12	15.51	1.54E-12	22.57	7.14E-22	23.29	2.26E-18
focal adhesion	27.08	4.84E-37	17.92	4.19E-48	20.12	1.72E-56	22.16	1.19E-52
actin cytoskeleton	25.1	1.69E-32	18.27	1.24E-48	18.99	1.32E-50	18.45	6.20E-39
toll-like receptor signaling	21.73	5.92E-09	19.37	3.48E-18	17.54	2.82E-15	21.3	1.54E-16
tight junction	21.91	4.41E-14	20.92	7.99E-32	18.47	1.11E-25	18.12	4.26E-20
chemokine signaling	22.24	3.35E-15	17.18	9.66E-25	17.5	6.09E-25	16.35	3.78E-18
TGFβ signaling	20.66	9.52E-11	18.41	1.07E-21	16.2	1.93E-17	16.56	1.38E-14
MAPK signaling	17.19	9.44E-22	18.83	9.21E-60	16.91	2.24E-49	16.97	4.24E-40
pathways in cancer	12.76	1.08E-18	13.98	1.90E-50	14.96	4.02E-55	16.11	4.11E-50

B **repressed**

pathway	0-30min ES	0-30min p-value	30-60min ES	30-60min p-value	60-90min ES	60-90min p-value	90-120min ES	90-120min p-value
axon guidance	32.31	2.91E-10	19.87	1.37E-07	20.61	1.07E-07	22.56	4.78E-10
cytokine receptor	25.19	2.73E-08	15.22	6.28E-06	15.78	5.08E-06	15.79	6.81E-07
apoptosis	22.04	7.42E-05	18.25	1.79E-05	15.78	2.71E-04	13.82	4.50E-04
ErbB signaling	12.02	2.50E-02	16.6	2.87E-05	20.08	1.20E-06	17.59	2.60E-06
pathways in cancer	16.63	2.99E-13	12.25	3.73E-12	14.29	8.51E-15	15.99	4.68E-20
purine metabolism	18.56	1.71E-05	12.81	1.02E-04	13.29	8.55E-05	13.57	1.21E-05
Jak-STAT signaling	16.95	2.09E-04	11.71	8.56E-04	9.71	7.98E-03	12.76	1.03E-04
ubiquitin proteolysis	7.67	1.50E-02	10.59	1.04E-05	9.61	8.85E-05	10.81	1.80E-06
MAPK signaling	8.96	5.44E-04	7.22	4.20E-04	10.69	3.99E-07	11.24	9.62E-09

C **repressed**

pathway	0-30min ES	0-30min p-value	30-60min ES	30-60min p-value	60-90min ES	60-90min p-value	90-120min ES	90-120min p-value
nucleotide excision repair	0	0	26.46	3.36E-05	21.95	7.55E-04	19.22	1.11E-03
cell cycle	0	0	13.53	3.10E-03	18.03	3.32E-08	21.06	9.50E-12
WNT signaling	0	0	8.7	2.58E-03	14.42	1.30E-06	14.21	2.16E-07
aminoacyl tRNA synthesis	0	0	0	0	28.69	2.42E-05	35.17	3.24E-08
p53 signaling	0	0	0	0	24.55	3.47E-07	33.78	5.64E-13

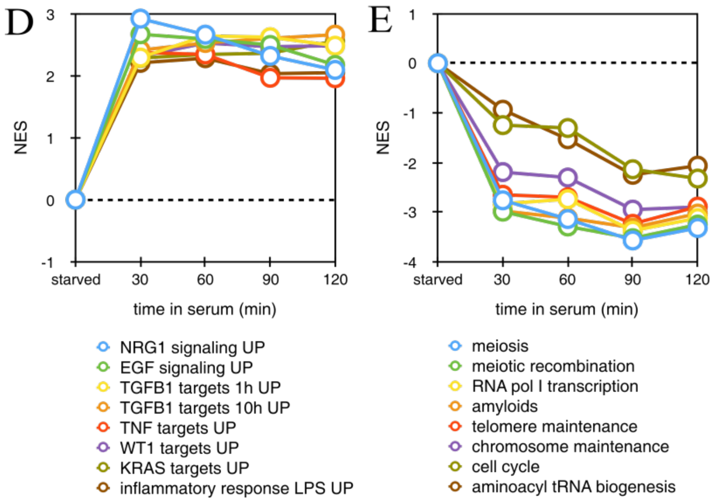


Figure 2. 6: Gene set enrichment analysis of serum response genes.

Enriched pathways identified by DAVID in induced (A) and repressed (B-C) gene sets during each serum stimulated labeling period. Enriched gene sets identified by GSEA in induced (D) and repressed (E) response gene groups during each labeling period. Normalized enrichment scores (NES) are displayed.

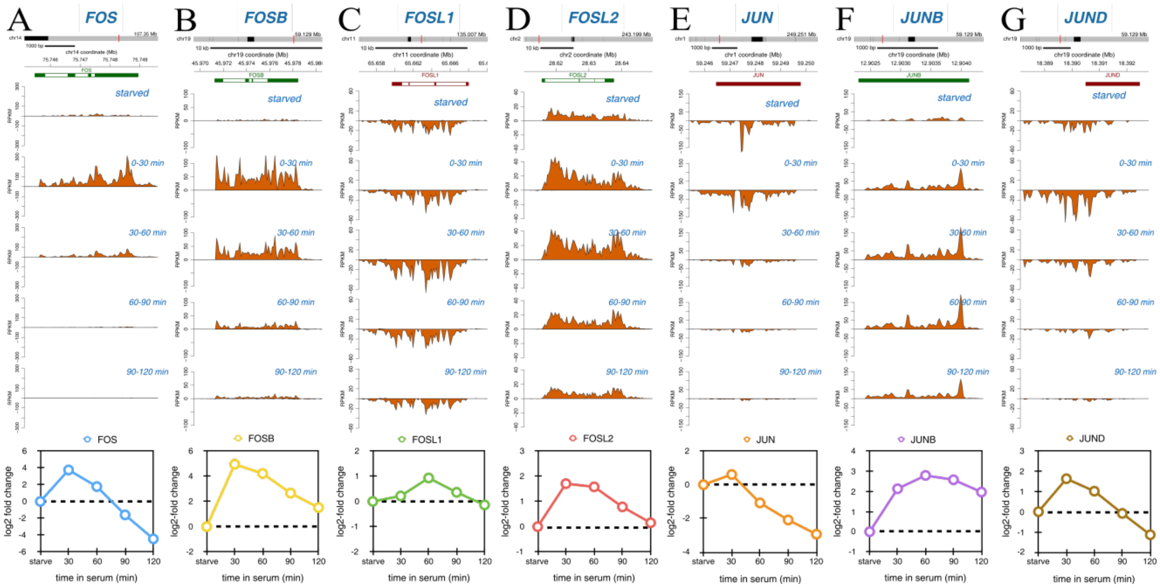


Figure 2. 7: Transcription of the AP-1 transcription factors family is transiently induced after serum stimulation.

Bru-seq traces for seven AP-1 transcription factor family genes during different periods following serum additions (top) and log2 fold changes compared to the serum starved sample (below).

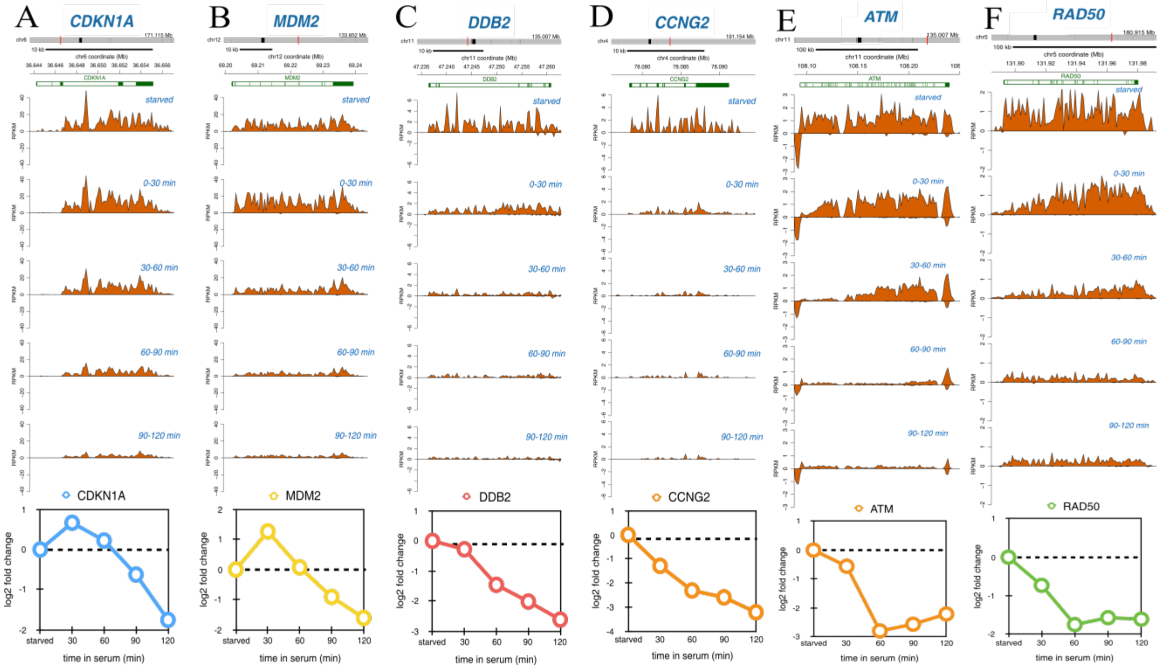


Figure 2. 8: p53 response and DNA damage response genes are downregulated following serum stimulation.

Bru-seq traces for p53 response genes (A-D) and DNA damage response genes (E-F) for starved cells and during different periods following serum addition.

short burst of transcription may stem from the rapid release of RNAPII promoter-proximal pausing, which is seen to occur near the FOS TSS[139,140]. Alternatively, the concurrent activation of inhibitory AP-1 factors may act to suppress subsequent rounds of transcription.

2.3.7 Downregulation of p53 response and DNA damage response genes following serum stimulation

In concordance with the pathway analysis, we observed transcriptional inhibition for a number of p53 response genes, either immediately or after a short delay (Fig. 2.8A-D). Serum starvation has been shown to increase protein expression of p53, with subsequent decrease after re-addition of serum[141]. Additionally, critical DNA damage response signaling genes were rapidly suppressed following serum stimulation. *ATM,* which encodes an important DNA damage response kinase, was downregulated during the first two hours following serum stimulation (Fig. 2.8E). Similarly, transcriptional repression was observed for *RAD50,* a gene encoding a protein component in the MRN complex, which may be required for ATM activation (Fig.2.8F). The functional implications related to the downregulation these DNA damage response genes during the serum response are not known.

2.4 Discussion

During a global cellular response, cells undergo widespread expression changes in order to modulate their function to adapt to changes in their environment. The ability to successfully follow the temporal chain of events occurring during a particular cellular response is critical for understanding the complex regulatory networks which orchestrate these gene expression changes. Categorizing response genes which display similar expression patterns can assist in identifying common regulatory mechanisms. Transcriptional regulation is a key step in establishing precise temporal expression of response genes. However, transcription elongation and RNA processing events create time delays for the generation of full-length mature mRNAs, and these delays are especially apparent for long genes. Therefore, by focusing on nascent transcriptome dynamics during a cellular response, we can more accurately assess early transcriptional events and their contributions to gene expression regulation.

In this study we presented the first genome-wide dataset of nascent transcription changes during serum activation. Because we focused on transcription occurring near TSSs, our characterization of the early serum response is a reflection of temporal patterns of productive initiation. This classification is distinct from previous studies that used steady-state RNA or mRNA detection and therefore categorized response genes based on transcript completion rather than initiation of transcription[130-132]. We identified 1417 induced and 636 repressed genes during the first two hours following serum addition, and these genes displayed diverse transcriptional patterns. We identified a set of genes that showed an immediate transcriptional response and genes that responded after a short delay. We were also able to categorize genes based on the behavior of their transcriptional over time, including whether the change in RNA production was sustained or transient. Genes that exhibit similar transcriptional patterns may be subject to the same mechanisms of transcriptional regulation. For example, at the level of transcription initiation, response timing might be related to enhancer status at the time of stimulation. Primed enhancers might regulate rapidly induced genes while de novo enhancer selection might result in a delayed response. Related to initiation, different sets or combinations of transcription factors or repressors may be responsible for immediate responses compared to delayed responses. Alternatively, regulation at the level of promoter-proximal pausing and release may also regulate response timing as well. Whether a transcriptional response is transient or maintained may be a factor of whether negative feedback mechanisms act to detect and limit RNA production levels. While this study did not explore specific mechanisms of transcriptional control, it provides a foundation for future study by establishing candidate gene sets that may be regulated together.

We hypothesize that these different transcriptional patterns are important for establishing temporal expression patterns related to functional activity. We performed DAVID and GSEA analyses to explore pathway and gene set enrichment patterns during the early serum response. We found that pathways which were highly enriched during the first labeling period, in either the group of upregulated or downregulated genes, tended to be enriched during the later labeling periods as well. Many of these enriched pathways were signaling pathways, such as the MAPK pathway, which is known to be activated by growth factor stimulation[142]. Attenuation of MAPK signaling and downstream target gene expression has been shown to depend on nascent transcription following activation[131]. Therefore, the enrichment of genes in these pathways may be related to transcriptional changes related to turning off the signaling pathway. MAPK activation results in transcriptional activation of AP-1 components *FOS* and *JUN*[143], which we were able to observe using Bru-seq. Induction of these genes is accompanied by the upregulation of alternate AP-1 components which may inhibit active AP-1 complexes, as has been shown for FOSL1 and JUNB[144,145]. The detection of genes induced or repressed at the same time may help to identify unknown regulators. However, the transcriptional response timing for these regulators is likely influenced by many additional factors, including RNA and protein abundance and degradation. Therefore, simultaneous assessment of transcriptional and translational regulation may be necessary to connect functional activity to temporal gene expression regulation.

Lastly, we observed the downregulation of p53 response and DNA damage response genes following serum stimulation. Protein levels of p53 are increased due to increased stabilization in serum starved cells, but these levels go back down following addition of serum[141,146]. The decrease of p53 target genes following serum addition is likely related to lower p53 protein levels compared to starved cells. Our Bru-seq data detects repression of p53 target *CDKN1A,* the gene that encodes the p21 protein, which is also seen to have increased levels in starved cells[146]. Elevated levels of p53 and p21 results in proliferation arrest. Surprisingly, we also observed decreased transcription of DNA damage response signaling genes *ATM* and *RAD50* following serum activation. Unlike p53, protein levels of ATM are not increased in starved cells, and it has been suggested that ATM signaling is inactive because cells are not progressing through the cell cycle[146]. Therefore, decreased expression would not be expected as cells are transitioning back into a proliferating state. It is possible that these changes in RNA production are accompanied by changes in RNA stability, translation, or protein stability, and may not have an overall effect on protein abundance. The functional implications related to the downregulation these genes DNA damage response genes following serum activation will need to be explored in more detail.

This study provides a novel view of the dynamic transcriptome as cells adjust to new environmental conditions after serum addition. Bru-seq revealed both known and previously unknown changes in transcription during the early serum response. The assessment of global transcriptional output provides clues as to how regulation of transcription is accomplished after cellular stimulation. We predict that nascent sequencing techniques such as Bru-seq will be critical to untangle the mechanisms behind temporal and dynamic expression changes for a wide range of cellular responses.

2.5 Materials and methods

2.5.1 Cell culture and serum stimulation

HF1, hTERT immortalized foreskin-derived human fibroblasts, were grown in MEM supplemented with 10% FBS, L-glutamine, vitamin mix, and antibiotics. Starved cells were grown in the same media minus FBS for 48 hrs. For serum stimulation experiments, FBS was added to the media of starved cells (final concentration 10%). Serum addition was followed by 30 min bromouridine labeling, either immediately or after a 30, 60, or 90 min incubation period. Bru-labeling was done for 5 samples: 1 starved sample and 4 serum stimulated samples for different serum incubation periods (Fig. 2.1).

2.5.2 Bru-seq analysis

Bru-seq was performed as previously described[105,147]. Briefly, bromouridine (Bru) (Aldrich) was added to the media of starved or serum stimulated cells at a final concentration of 2 mM and incubated at 37°C for 30 min. Total RNA was isolated using TRIzol reagent (Invitrogen), and Bru-labeled RNA was isolated by incubation with anti-BrdU antibodies (BD Biosciences) conjugated to magnetic Dynabeads (Invitrogen) under gentle agitation at room temperature for 1 h. cDNA libraries were prepared from the isolated 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 were carried out as previously described[105,147].

2.5.3 Gene expression and serum response analysis

RPKM (reads per kilobase per million mapped reads) values were calculated for individual genes over 300 bp for starved cells and for each serum stimulated sample. For genes 30 kb and under, RPKM was calculated using read counts from the entire gene. For genes over 30 kb, an RPKM value was calculated using read counts from the first 30 kb downstream of the TSS. Genes were classified as expressed if they had an RPKM value greater than 0.5 in starved cells or in at least one serum stimulated sample.

To identify response genes, an inter-sample comparison analysis was done to obtain RPKM fold change values for each gene in a given serum stimulated sample compared to the starved sample[105]. Genes with a greater than 2-fold change in any serum stimulated sample compared to starved cells were categorized as serum-response genes. Genes with greater than a 2-fold increase were classified as induced and genes with greater than a 2-fold decrease in RPKM values were classified as repressed. Repressed genes were required to be expressed (RPKM value > 0.5) during the starved condition.

For the generation of the heatmaps, induced and repressed response genes were clustered based on when the 5' end of the gene reached a greater than 2-fold change compared to starved cells. Genes were also clustered based on whether the expression change was sustained or transient. Sustained response genes maintained a greater than 2-fold change in the labeling periods following the initial response. Transient response genes had expression levels that returned to starved levels (less than 2-fold change) during the labeling periods following the initial response.

2.5.4 Pathway and gene set enrichment analysis

We used DAVID Bioinformatics Resource 6.7[134,135] to perform gene set enrichment analysis on the serum-response genes. The background gene set used for the analysis contained all genes that were expressed above 0.5 RPKM in our cells. We performed a functional annotation analysis using induced and repressed response gene sets from each labeling period. Here we present enriched pathways with a p-value < 0.05. We also performed gene set enrichment analysis (GSEA)[137]. All genes expressed >0.5 RPKM were rank-ordered according to log2-fold changes compared to the gene expression in serum-starved cells.

2.5.5 Author information

KSK and MTP contributed to experiments. KSK, BM, KB, and ML contributed to data analysis. KSK and ML contributed to figure production and manuscript writing.

CHAPTER III

Gene length as a biological timer to establish temporal transcriptional regulation

3.1 Abstract

Transcriptional timing is inherently influenced by gene length, thus providing a mechanism for temporal regulation of gene expression. While size has been shown to be important for the expression timing of specific genes during early development, whether it plays a similar role during global gene expression programs has not been extensively explored. In this study, we investigate the role of gene length during the early transcriptional response of human fibroblasts to serum stimulation. Using the nascent sequencing technique Bru-seq, we identified immediate genome-wide transcriptional changes following serum stimulation. Immediately induced genes, including transcription factors, displayed a wide range of sizes which results in staggered production of their gene products. Immediately repressed genes also exhibited a wide range of sizes, but the median value of downregulated genes was shorter than that of other genes. Corresponding mouse orthologs of these serum response genes also vary in size, and relative gene size appears to be evolutionarily conserved. Additionally, we also demonstrated that RNA polymerase II accelerates as it transcribes large genes, but that this was independent of whether the gene was induced or not.

Our Bru-seq results provide a comprehensive profile of nascent transcription during the immediate serum response. Variations in gene size allow for a large group of genes to be simultaneously activated but still be expressed at different times following serum stimulation. The sizes of immediately induced transcription factor genes varied dramatically, setting up a cascade mechanism for delayed induction of downstream genes. The order in which genes are expressed following activation of a particular signal transduction pathway may be important for efficient protein function. This gene expression order is likely to be conserved in other species because relative gene size has been maintained during evolution. While elongation of extremely long genes can take

several hours, acceleration of RNA polymerase II may act to reduce time delays. We demonstrate the role of gene length in establishing genome-wide expression timing during the serum response, and predict that gene length also influences the timing of other gene expression programs.

3.2 Introduction

Human genes come in a wide variety of lengths; protein-coding genes range from less than a hundred to over two million base pairs long[148]. One consequence of this extensive range in gene size is that the time needed to complete transcription elongation of a full-length transcript is also highly variable among genes. Hypothetically, at a constant estimated transcription elongation rate of 1.5 kb/min[68,69,149,150], a 100 bp gene would take a few seconds to complete elongation while a 2 Mb gene at the same rate would take over 20 hours to complete. Long genes tend to be mostly comprised of intronic sequences that are transcribed but then excised by the spliceosome during RNA splicing. Additionally, introns tend to be longer in humans than in other vertebrate species[151], and longer in tissue-specific genes compared to housekeeping genes[152]. Furthermore, long genes can promote genomic instability and are associated with the formation of copy number variations (CNVs) and fragile sites[153]. Why have long genes survived evolutionary pressures for shortening despite the time and energy burden and the threat to genomic integrity that they present? While introns undoubtedly play critical roles in RNA regulation, processing, and isoform diversity[154,155], their contribution to gene size variability suggests that intron length may also serve an important biological function in regulating the temporal expression of genes induced by a cell stimulus or as part of a stress response.

Early observations of large introns in *Drosophila* developmental genes led to the "intron delay hypothesis" which postulates that intron length may play a role in gene expression timing[156]. Indeed, it has been demonstrated that during *Drosophila* early development, the gap gene *kni* is expressed while its cognate gene *knrl*, which has more intronic sequence, does not have enough time to be expressed due to rapid cell divisions disrupting its transcription[157]. More recently, it was shown that intronic length is important during somite segmentation of mouse embryos[158]. During this time, the *Hes7* gene is cyclically expressed as a result of regulation through a negative feedback loop involving the Hes7 protein. Removal of introns from the *Hes7* gene led to earlier expression, abolishment of oscillatory expression, and developmental defects. While intron length is implicated in the regulation of expression timing for certain genes during development, whether it is important for coordinating gene expression programs outside of development is unclear.

Gene expression programs allow cells to react to specific changes in their environment through temporally coordinated transcription of response genes. Various signaling cascades activate preexisting transcription factors, which induce primary response genes. Immediate-early genes induced as part of the primary response often encode transcription factors or signaling factors and are rapidly and transiently expressed[133]. The target genes of these transcription factors are part of the secondary response in which gene induction requires *de novo* protein synthesis and therefore necessitates a delay before gene expression. Initiation is classically viewed as a key regulatory point for gene expression timing, however other transcriptional events, such as escape from promoter proximal pausing and transcription elongation, are also likely to influence temporal expression patterns during the activation of gene programs[159,160].

The early serum response provides an excellent model for studying immediate-early gene expression programs. Fibroblasts grown in serum-free media enter a G_0 state of proliferation arrest, and subsequent addition of serum activates global gene expression changes as cells begin to progress through the cell cycle again[129]. While the genome-wide early serum response has been previously explored, earlier studies have focused on expression changes based on total mRNA levels[130,131]. Although mRNA expression provides information about the steady-state of completed transcripts, it does not indicate timing of upstream transcriptional events, such as induction or repression of transcription initiation. Some studies have explored nascent transcription of early response genes using qRT-PCR[132,159], but this method is limited to the transcriptional assessment of a handful of selected genes. We used the nascent RNA sequencing technique Bru-seq to assess immediate genome-wide changes in transcription following serum stimulation of serum-starved human fibroblasts[147]. Because Bru-seq allows us to assess transcription initiation before the generation of a final mRNA product, our study provides a comprehensive profile of the immediate transcriptional response to serum. We detected a set of immediately induced genes, many which were previously identified, as well as a novel list of genes immediately repressed in response to serum stimulation. Our results highlight the gene size variability of immediate response genes and point to a role for gene length in temporal global expression timing.

3.3 Results

3.3.1 Identification of immediate serum response genes using Bru-seq

To investigate the immediate transcriptional response to serum, we compared nascent RNA expression in starved and serum activated normal human fibroblasts using Bru-seq[147]. Cells were grown in serum-free media for 48 hours, and then serum was added back to the media (or not for the starved control) and nascent RNA was immediately labeled with bromouridine (Bru) for 30 minutes (Fig. 3.1A). The Bru-RNA was isolated, used to prepare cDNA sequencing libraries, and mapped to the reference genome. Results from two biological experiments were highly correlated (Fig. 3.2). Among the genes that were immediately upregulated, we observed many known immediate-early genes such as the transcription factor gene *FOS* (Fig. 3.1B). This was evident through an increase in reads across the entire gene. In long induced genes such as signaling gene *PDE7B* (~344 kb), the increase in reads was present at the beginning of the gene but only extended partially into the body of the gene (Fig. 3.1C). We previously measured a median RNA polymerase II (RNAPII) elongation rate of approximately 1.4kb/min in our fibroblast cell line[68]. Consistent with this rate, the 30 minute labeling period did not allow enough time for newly initiated *PDE7B* transcripts to complete transcription. Bru-seq also allowed us to identify genes that were transcriptionally repressed immediately after serum addition, such as signaling gene *TRIB2* (Fig. 3.1D). For long repressed genes, such as signaling gene *GNG2* (109 kb), we observed a decrease in reads at the beginning of the gene and a receding wave of reads towards the 3’end of the gene (Fig. 3.1E). This receding wave corresponds to transcripts initiated prior to serum addition that continued to elongate towards the 3’end of the gene during the labeling period.

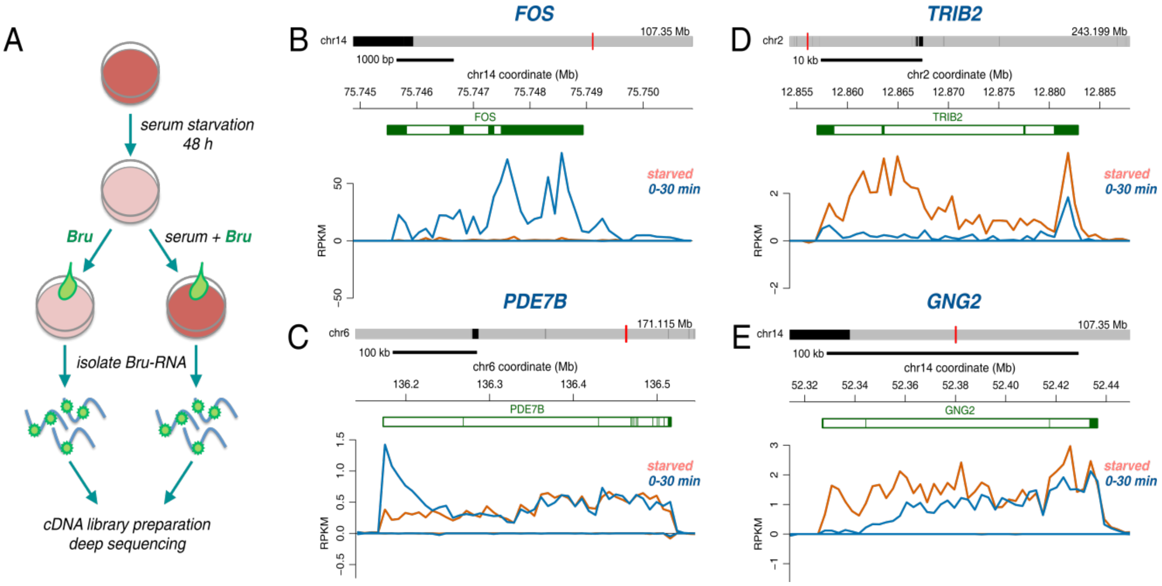


Figure 3. 1: Bru-seq to capture immediate transcriptional changes during the serum response.

(A) Experimental outline. Bromouridine (Bru) labeling of nascent RNA was performed for 30 minutes on starved or serum stimulated human fibroblasts. Nascent RNA sequencing reads for *FOS* (B), *PDE7B* (C), *TRIB2* (D), and *GNG2* (E) during starved conditions (orange) and following serum addition (blue). Reads are mapped to the reference genome, with annotated genes shown at the top in green.

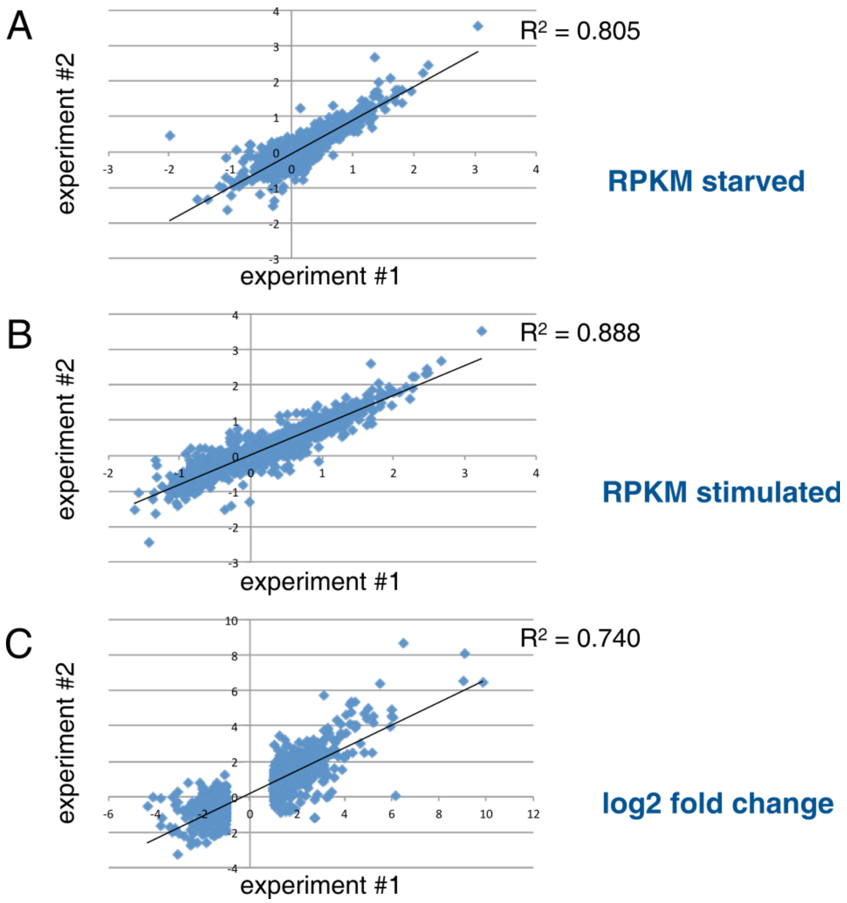


Figure 3. 2: Correlations between the two independent biological experiments.

The serum stimulation was performed in HF1 human fibroblasts in two replicate experiments. RPKM values are from 5931 genes expressed >0.5 RPKM in both experiments. (A) Comparisons of RPKM values between the two starved samples. (B) Comparisons of RPKM values between the two serum stimulated samples. (C) Comparison of the log2 fold change induced by serum between the two experiments.

For our genome-wide analysis, we calculated nascent RNA expression levels for all genes in starved cells and during the 30 minute serum stimulation period. For genes over 30kb, we calculated expression level based on the first 30kb since we predicted that response genes would exhibit read changes within this region but not at the end of the genes. Out of the 6958 genes fitting our expression criteria of being at least 300bp long and expressed above 0.5 RPKM, we identified 873 significantly upregulated and 209 significantly downregulated genes (adjusted p-values <0.05, n=2) during the first 30 min of serum stimulation. Our results demonstrate that serum stimulation of starved human fibroblasts results in rapid global transcriptional changes.

Previously defined groups of immediate-early genes based on increases in steady state mRNA levels tend to be dominated by small genes[132,133]. This is partially due to timing constraints on elongation since these techniques are biased towards detection of full-length transcripts. Bru-seq allows for the assessment of instantaneous changes in transcription by analyzing reads immediately downstream of transcription start sites (TSSs). A comprehensive analysis of immediate transcriptional changes is not afforded by microarray or conventional RNA-seq because changes in steady state RNA pools are delayed due to the noise of pre-existing RNA levels. Furthermore, if RNA is isolated through poly(A) selection, changes will only be detected after the completion of full length, processed transcripts. Thus, Bru-seq is a powerful tool for detecting immediate changes in levels of transcription initiation, and can be used to explore the relationship of transcriptional changes with the temporal expression patterns of full length RNA.

3.3.2 Identification of primary transcription factors responding to serum stimulation

Because our group of upregulated genes is large and inclusive, we expected to be able to identify potential transcription factors that may be responsible for mediating the induction of large groups of immediate serum response genes. We performed a gene set enrichment analysis (GSEA) using transcription factor targets gene sets, which are characterized based on the presence of specific transcription factor binding motifs. We ranked all genes according to fold changes in expression in serum stimulated cells compared to starved cells. This ranked gene list was used to calculate normalized enrichment scores (NESs). Unsurprisingly, the binding motif for serum response factor

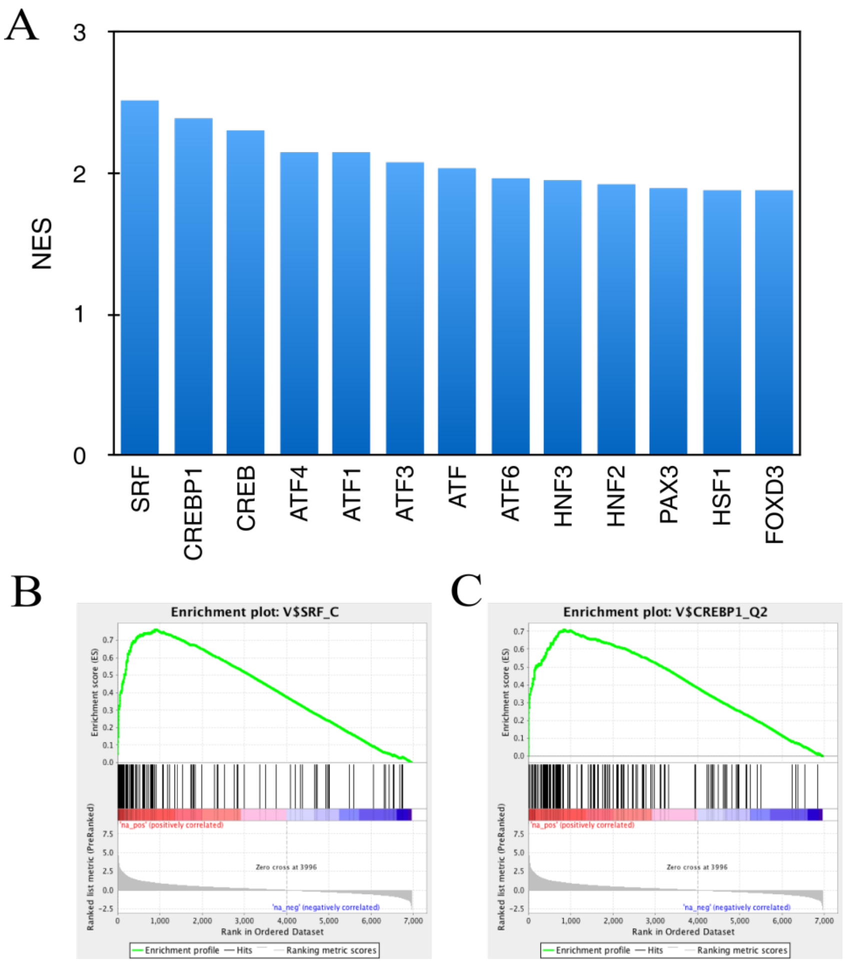


Figure 3. 3: Transcription factor binding motif enrichment in immediately induced serum response genes.

Normalized enrichment scores calculated by GSEA for highest scoring transcription factor binding sites identified in induced serum response genes (A). Enrichment plots for SRF (B) and CREBP1 (ATF2) (C).

(SRF) had the highest enrichment score, indicating its presence at high frequency among highly induced genes (Fig. 3.3A,B). Other binding motifs with high enrichment scores included those for the ATF/CREB family of transcription factors (Fig. 3.3A,C).

3.3.3 Early response genes complete transcription at different times due to a wide range of lengths

Our data indicated that immediate response genes exhibit a wide distribution of genes sizes, as do all genes that are expressed in either the starved or serum stimulated cells (Fig. 3.4A). Therefore, even though the response genes are induced or repressed around the same time, differences in gene length result in various temporal expression patterns. Statistical comparisons of the size distributions for these groups of genes revealed that induced genes tend to be larger with a median value of ~37.0 kb and repressed genes tend to be shorter with a median value of ~15.7 kb, compared to all expressed genes with a median value of ~29.5 kb. For comparison, housekeeping genes expressed in our cells were significantly shorter, with a median value of ~9.4 kb. The variance of induced genes was slightly but significantly higher than that of all expressed and repressed genes. These differences in the size distributions reveal that the expression timing of induced serum response genes is influenced by elongation delays more than repressed serum response genes. A similar pattern, with longer induced genes larger than repressed genes, was observed for genes that display transcriptional changes during the 60 to 90 minute period following TNF stimulation (data not shown). This suggests that early induced genes tend to be larger than early repressed genes in general, and that this is not a unique pattern for serum response genes.

Among the immediate response genes identified by Bru-seq was a large group of transcription factor genes. Transcription factor genes made up approximately 13% of the upregulated response genes (111/873), approximately 10% of the downregulated response genes (21/209). These transcription factor genes were also various sizes and estimated to complete transcription at different times (Fig. 3.4A,B). The median length of the induced transcription factors was ~28.8kb, which was similar to the median of all expressed transcription factors (~27.3kb) and expressed genes. Repressed transcription

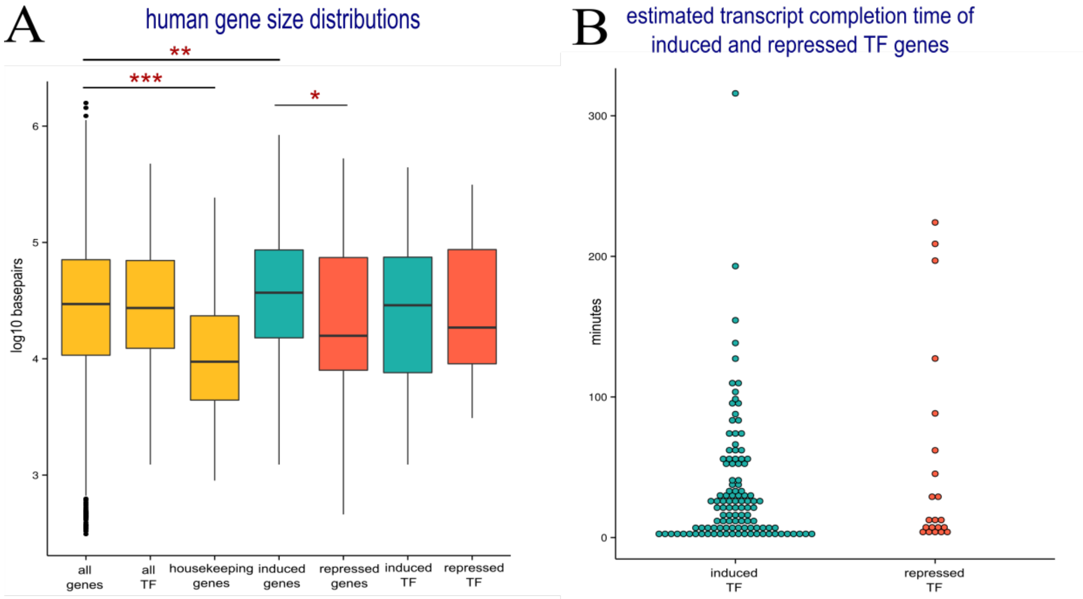


Figure 3. 4: Serum response genes exhibit a broad range of gene sizes.

(A) Boxplots displaying the distributions of gene sizes for various gene sets: all genes expressed in either starved or serum stimulated cells (6958 genes), transcription factor genes expressed in either starved or serum stimulated cells (488 genes), housekeeping genes expressed in either starved or serum stimulated cells (336 genes), genes induced after serum stimulation (873 genes), genes repressed after serum stimulation (210 genes), transcription factor genes induced after serum stimulation (111 genes), and transcription factor genes repressed after serum stimulation (21 genes). Asterisks indicate statistical significant differences between groups using Mann-Whitney-U test with p-values *<2x10^{-5}, **< 2x10^{-7} and ***< 3x10^{-16}. (B) Induced and repressed transcription factor genes displayed according to estimated time needed to complete transcription.

factors were shorter, with a median value of ~ 18.6kb, however the number of genes in this category was relatively small.

Based on size, small genes such as transcription factor genes *FOSB*, *NR4A1*, and *NR4A2* are expected to produce full-length transcripts within about 8 minutes (Fig. 3.5A-C). Medium-sized genes such as signaling genes *CDK7*, *REL*, and *LIMA1* are expected to produce full-length transcripts after about 30 minutes (Fig 3.5D-F). Because increased transcription levels are not seen at the 3'end of these genes during the initial 30 minute time period, this suggests that the induction of these genes occurred at various times within the labeling period after serum addition. Large genes such as *PDLIM5*, *BTAF1*, and *UBR4*, which have various functions, are expected to produce full-length transcripts after 1-3 hours (Fig. 3.5G-I). Indeed, we see increased reads at the ends of these longer transcripts when labeling is done during later periods following serum stimulation (data not shown).

Thus, though these transcription factors are all induced within the first 30 minutes of serum addition, variations in gene size will stagger their production, thereby setting up a temporal cascade of gene regulation. We ordered the induced transcription factors according to size and estimated transcriptional completion timing based on length and the median elongation rate (Fig. 3.4B). While we estimate that the majority of the immediately induced transcription factors complete transcription within the first hour after serum stimulation, there are several transcription factors (21/111) that are likely to complete transcription later, some not until two hours following serum addition. Following translation, these transcription factors likely induce another set of target response genes, and expression timing of these target genes would depend on gene size as well. This sets up a complex regulatory mechanism by which successive series of gene induction events set up precise temporal expression patterns as a function of gene size.

MicroRNAs (miRNAs) are small non-coding RNAs that can target specific mRNAs for translational inhibition or degradation[161]. Annotation of miRNA genes has been difficult because rapid processing of miRNA precursors into mature miRNAs does not allow traditional RNA-seq techniques to capture miRNA primary transcripts. We

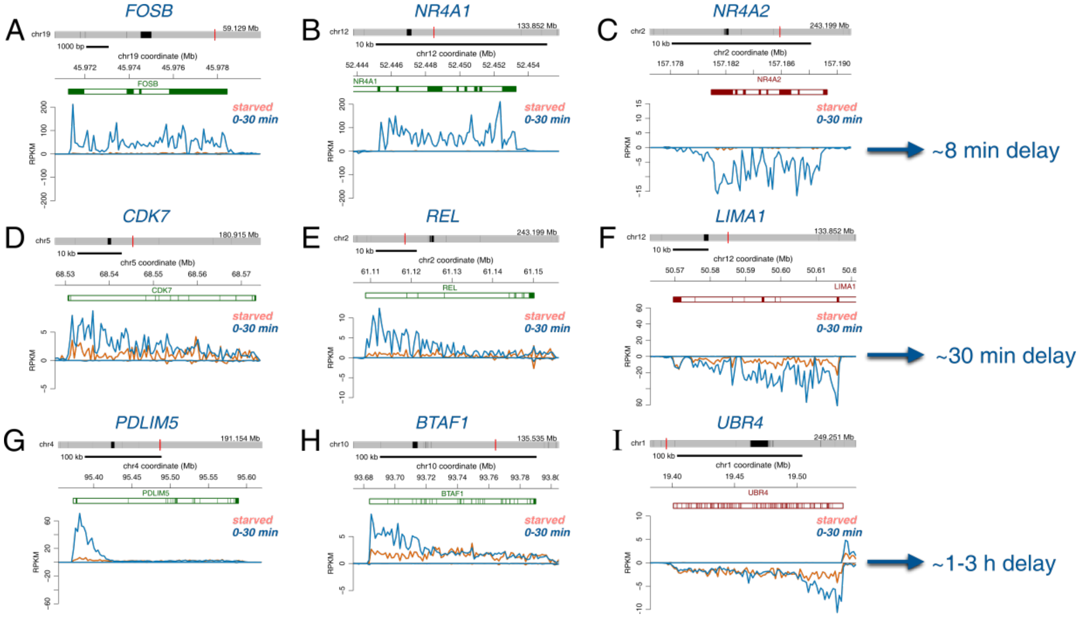


Figure 3. 5: Estimated transcriptional delays as a result of gene length.

Bru-seq traces for (A), *FOSB*, (B) *NR4A1*, (C) *NR4A2*, (D) *CDK7*, (E) *REL*, (F) *LIMA1*, (G) *PDLIM5*, (H) *BTAF1*, and (I) *UBR4* during starved conditions (orange trace) and following serum stimulation (blue trace), along with estimated times required for transcription completion based on length and an elongation rate of 1.4 kb/min on the right.

have used Bru-seq to identify and annotate genome-wide miRNA host genes in multiple cell lines, and found that they dramatically vary in size (unpublished data). Here we found several miRNAs genes that are transcriptionally upregulated in response to serum. *MIR143HG* is an example of an annotated miRNA host gene which is ~26kb and gives rise to *MIR143.* We detected rapid induction of the miRNA gene following serum stimulation (Fig. 3.6A). For *MIR30B* and *MIR30D*, there is no current host gene annotation, but we observed a ~48kb transcript which started approximately 27kb upstream of these miRNAs and was also upregulated in response to serum addition (Fig. 3.6B). We identified other miRNAs, such as *MIR21* and *MIR3193*, that were located downstream of highly induced genes (Fig. 3.6C,D). We hypothesize that these miRNAs are regulated by transcription run-on past the termination sites of the upstream genes. Because the upstream gene transcripts are short (less than 3kb), transcription of both the gene and miRNA are expected to be completed quickly. We also observed two miRNA clusters on chromosome 14 whose transcription appeared to be regulated by the transcription of the upstream gene *MEG3*, an annotated lncRNA (Fig. 3.6E). The first miRNA cluster consists of 10 miRNAs and is ~43kb downstream of the *MEG3* TSS, and the second cluster consists of 43 miRNAs and is located ~196kb downstream of the TSS. Therefore, we predict that the transcription of the first cluster would be completed about 40 minutes after the induction of *MEG3*, but the second cluster would take around 2 hours to produce mature miRNAs. When labeling is done during the 90-120 minute period following serum stimulation, we see increased reads at the second miRNA cluster (data not shown). Timing delays in the production of miRNAs set up temporal regulatory cascades similar to the ones established by transcription factors in which gene size influences when these transcriptional regulators can act upon their targets to control gene expression.

3.3.4 Relative gene length is evolutionarily conserved

If gene length plays an important role in coordinating expression timing, gene size would be predicted to be evolutionarily conserved. Previous studies have shown that genes tend to be larger in humans compared to other animals, largely due to intron lengthening [151,162,163]. In comparisons of human and mouse orthologous introns, 70% of

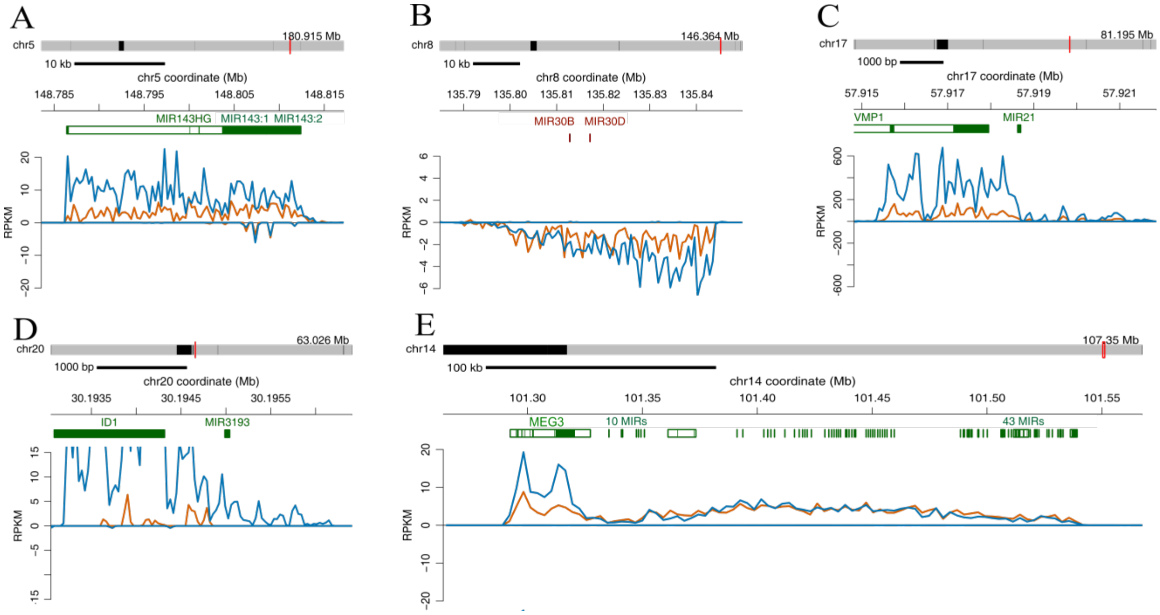


Figure 3. 6: Transcriptional regulation of miRNA host genes of various sizes during serum stimulation

Bru-seq traces for miRNA host genes during starved conditions (orange trace) and following serum addition (blue trace).

human introns were larger; however there was still a strong correlation between the lengths of the pairs[163]. We examined the relationship between total gene length of human and mouse orthologs, and similarly observed a strong correlation (R^2 =0.84) (Fig. 3.7A). We then focused on the immediate serum-induced transcription factors and compared their length to orthologous mouse genes. We observed that when transcription factor genes are ordered according to size, their corresponding mouse orthologs have a similar ranked order (Fig. 3.7B). These results suggest that the maintenance of relative gene size may be important for proper expression timing.

3.3.5 Functional annotation analysis based on response gene size

Because gene length appears to be evolutionary conserved, this suggests that it may play a functional role in the regulation of gene expression during cellular responses. Gene length may be important for ordered translation of mRNAs by ribosomes and organized production of proteins involved in different response pathways. To examine whether induced genes of different sizes were enriched for specific biological pathways, we grouped the 873 serum-induced genes based on size and performed DAVID functional annotation analysis (Fig. 3.8A-E). The induced gene groups were compared to a background of all expressed genes within the same size category in order to eliminate biases due to pathways which are overrepresented in certain size categories. Genes involved in “focal adhesion” and “actin cytoskeleton” were enriched in all size categories, and “MAPK signaling pathway” genes were enriched in 4 out of the 5 size groups. Genes involved in the “regulation of transcription” were enriched in the three smallest size classes. For the 209 genes repressed by serum stimulation, we created two size groups: smaller or larger than 15 kb. Similar to the induced gene groups, we found enrichment of genes involved in “transcription”, “actin cytoskeleton”, and “focal adhesion”, however the enrichment of these was only found in one size group (Fig. 3.8F,G).

3.3.6 Transcription accelerates towards the 3’-end of long genes

Extremely long genes are estimated to take several hours to complete transcription. However, it has been suggested that elongation rates accelerate as RNAPII travels across the gene body, and this could act to reduce transcriptional delays[67,69]. We

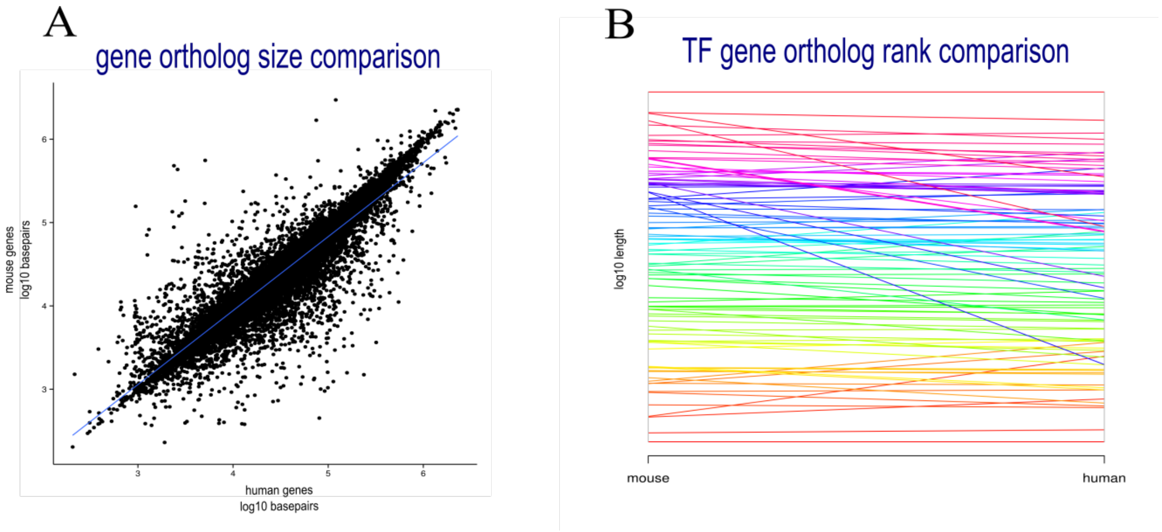


Figure 3. 7: Human genes and their mouse orthologs sizes are correlated.

(A) Correlation plot of sizes for all human protein-coding genes and their mouse orthologs (15845 genes) (R^2 =0.84). Only orthologs with one-to-one homology were assessed. (B) Serum induced TF genes were ranked according to size (right) and compared to the ranking of corresponding mouse orthologs (left). Each gene-ortholog comparison is represented by a single line.

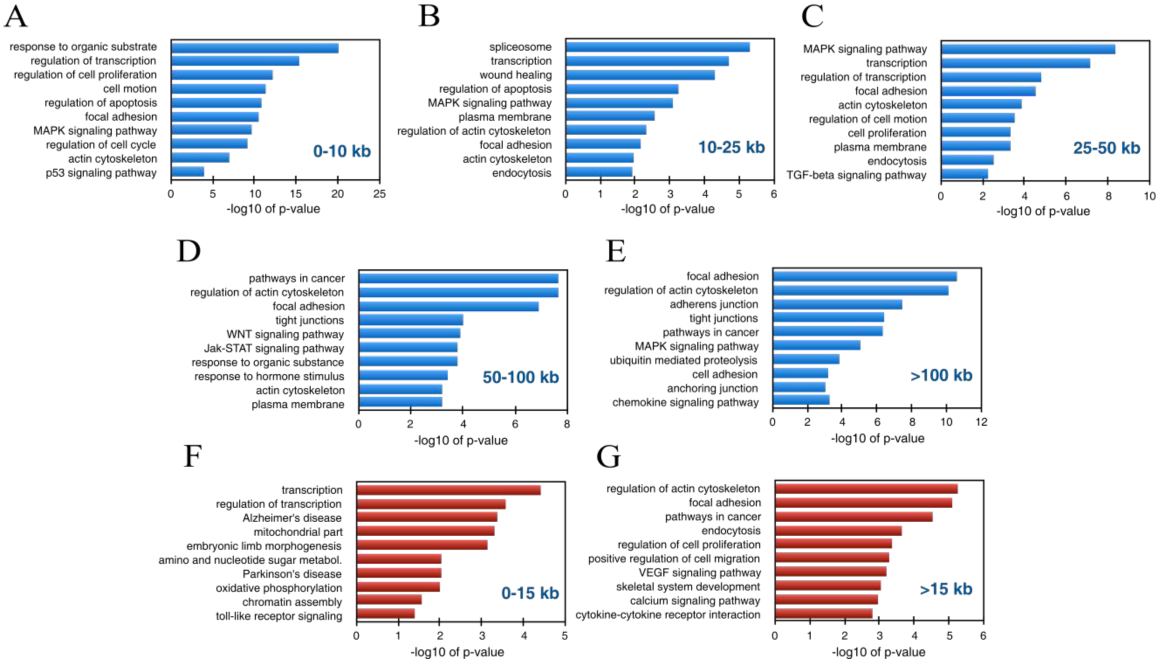


Figure 3. 8: Functional annotation analysis for immediate serum response genes.

Enriched pathways identified by DAVID in induced (A-E) and repressed (F-G) gene sets grouped based on gene size. The data is expressed as -log10 p-values.

previously assessed genome-wide transcription elongation rates in our fibroblast cell line by using BruDRB-seq to measure distances traveled by RNAPII during the first 10 minutes following the release of a DRB-induced arrest at promoter-proximal sites[68].

During the serum response, the synchronization of transcriptional responses can also be used to explore elongation rates. We found that the 30 min Bru-labeling interval done immediately after serum addition consisted of variable distances travelled by RNAPII, probably due to response genes starting transcription at different times within the labeling period. Therefore, we examined the distances traveled by the transcription wave of long genes (>200 kb) during the later labeling periods: 30-60, 60-90, and 90-120 minutes following serum stimulation. Transcription rates for each gene were calculated using the genomic distance between the leading edge of transcription waves from two sequential time intervals divided by the 30 minute labeling time. The median distance covered by the transcriptional wave across 21 induced genes was 60kb, 90kb, and 100kb during the 30-60, 60-90, and 90-120 minute labeling periods, respectively. This translates into an average elongation rate of 1.97 kb/min, 2.79 kb/min, and 3.32 kb/min for the respective time intervals, and supports the idea that RNAPII accelerates as it travels across large genes (Fig. 3.9A-C,G). Long genes repressed by serum exhibited a retreating wave of transcription which was used to measure elongation rates a similar way. In 21 repressed genes, RNAPII was estimated to travel a median distance of 70kb, 90kb, and 100kb during the 30-60, 60-90, and 90-120 minute labeling periods, respectively. The corresponding elongation rates were 2.33 kb/min, 3.19 kb/min, and 3.25 kb/min for the respective time intervals (Fig. 3.9D-F,H). Because these distances are similar to those measured using induced genes, elongation rates were similar before and after serum stimulation. These findings suggest that transcription elongation accelerates as it traverses large genes, but that these increased elongation rates are not the result of gene induction caused by serum stimulation but rather an inherent feature of transcription over long distances.

3.4 Discussion

Though early gene responses have been well studied, dissecting the different mechanisms contributing to temporal gene expression has been difficult due to large

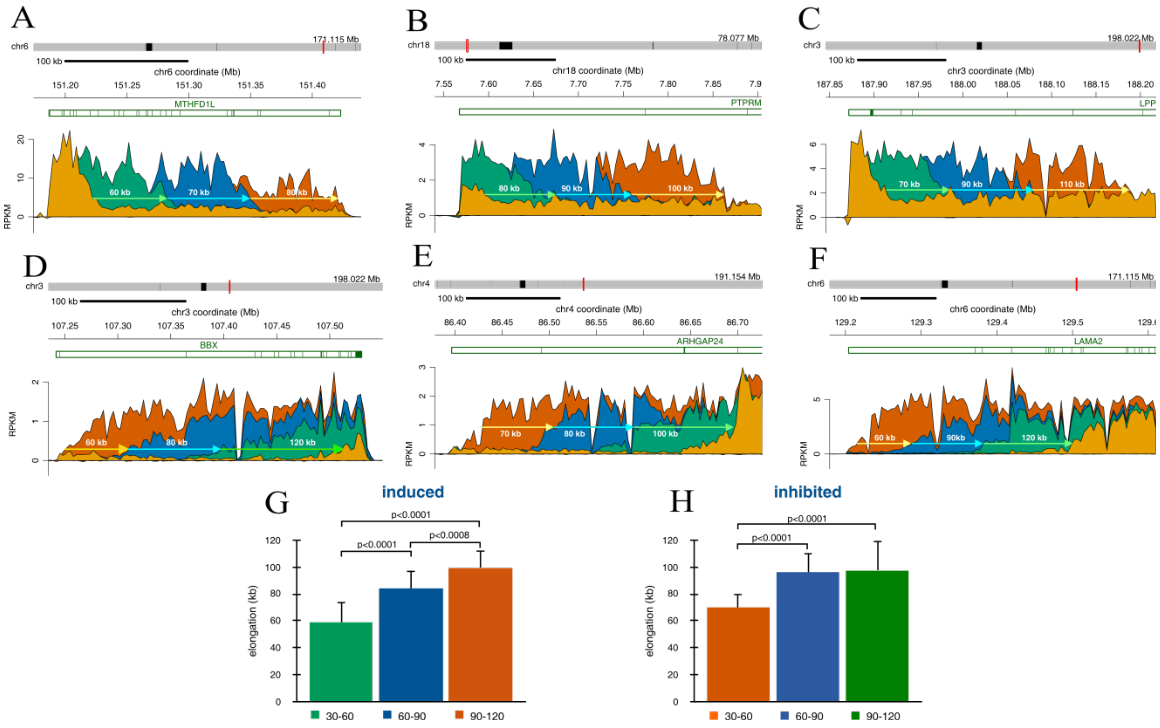


Figure 3. 9: RNAPII accelerates during elongation across large genes.

Bru-seq traces for large induced genes (A-C) during the different labeling periods (0-30min yellow, 30-60min green, 60-90min blue, 90-120min orange). Bru-seq traces for large repressed genes (D-F) during the different labeling periods (0-30min orange, 30-60min blue, 60-90min green, 90-120min yellow). Estimated distances of each transcription wave are indicated by the arrow. The mean of the estimated elongation rates of 21 genes with standard deviations are displayed for induced (G) and repressed (H) genes for the different labeling periods. The p-values are indicated above each comparison.

amounts of response genes and numerous points of regulation. In order to explore common regulatory mechanisms, response genes are often categorized according to expression timing. These gene groups have been determined based on when changes in the total RNA pool are detected, and therefore the expression timing reflects completion of transcription and RNA processing. However, productive initiation is one of the earliest steps during a gene response, and we feel that the timing of initial induction together with gene length should be considered when describing response gene kinetics. Additionally, information on transcriptional timing can help to distinguish whether expression delays are occurring at the level of transcription or post-transcriptionally. In this study, we used Bru-seq to conduct a genome-wide profile of nascent transcription during the early serum response. Our results demonstrate that transcriptional timing information can provide valuable insight into the mechanisms contributing to temporal gene expression regulation during cellular responses.

Initiation is often seen as a key regulatory point for gene expression, and recruitment of RNAPII and promoter proximal pausing likely contribute to variations in temporal expression patterns during early cellular responses such as the serum response[159,160]. However, gene length and transcription elongation also act as regulatory mechanisms which influence expression timing. Immediate-early expressed genes have been typified by their short length and rapid induction. Our genome-wide profile of nascent transcription indicates that for the serum response, these common traits are not due to selective initiation of small genes but instead are a result of elongation constraints. We observed hundreds of genes being induced immediately, yet due to the broad range of gene lengths and time needed to complete elongation, various temporal expression patterns are established. Immediately induced transcription factors similarly vary in size, and therefore gene length also acts as mechanism that regulates the induction timing of secondary response target genes (Fig. 3.10). Similarly, miRNA host genes display a range of sizes, thereby regulating transcriptional repression in a complex temporal manner (Fig. 3.10). Moreover, some miRNAs are downstream of highly induced genes, and may act through negative feedback loops. Induced genes tend to be larger, demonstrating how this set of genes will experience increased time delays after serum

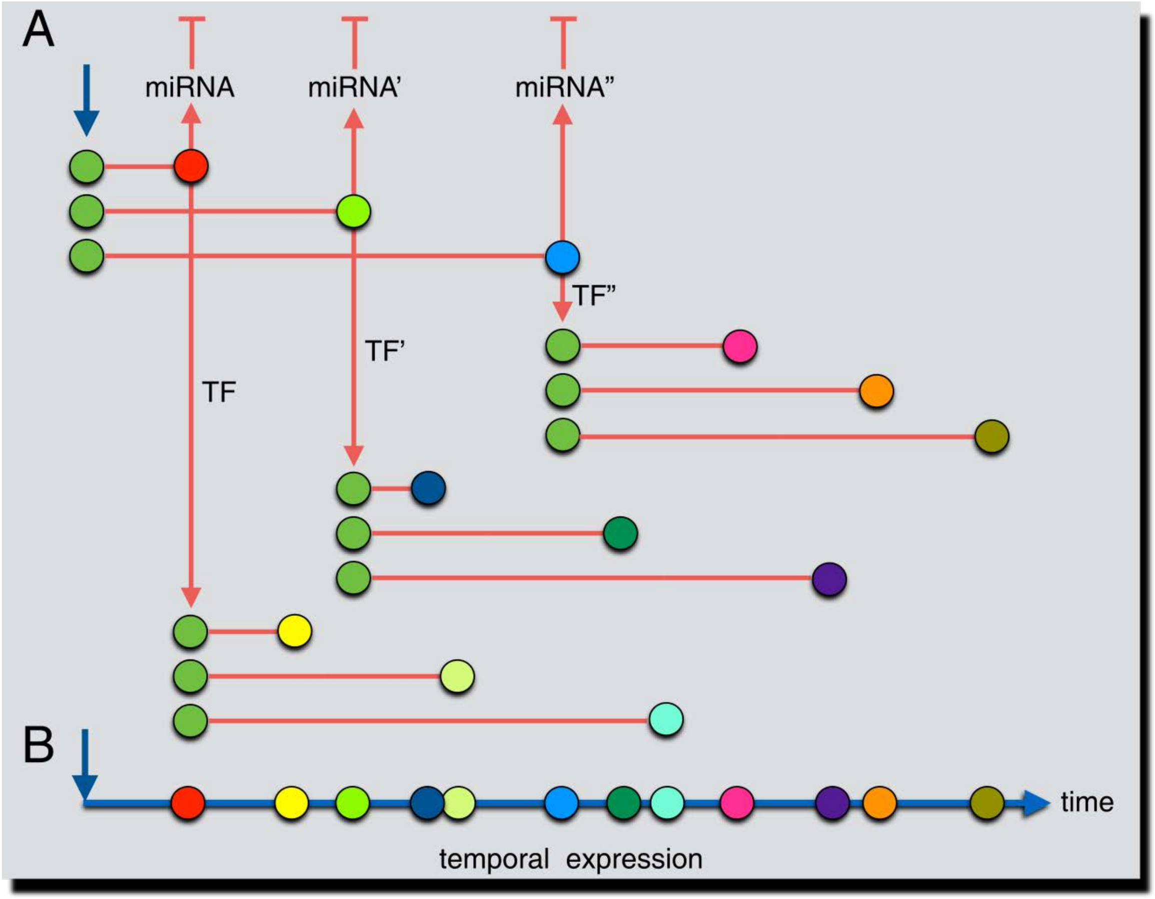


Figure 3. 10: Model for the role of gene length in establishing temporal expression patterns following serum stimulation.

(A) Transcription induction occurs simultaneously for several genes (first set of green dots on left), but gene length influences the completion timing of the transcript (red lines and colored dots). These immediately induced transcription factors (TF) and miRNAs then go on to activate or inhibit their own gene targets (green dots), whose expression timing is also influenced by gene length. (B) Complex, staggered expression timing established by gene length.

stimulation. In contrast, repressed genes tend to be shorter and will not be equally affected by transcriptional delays. If the precise timing of expression is not important, this may result in a selection for shorter gene length. Housekeeping genes, which are typically constitutively expressed in order to maintain basic cellular functions, are seen to be shorter overall. Therefore, the shorter size of repressed genes may represent the need to be turned off rapidly, or could indicate that the timing of downregulation for these genes is not critical during the serum response.

The role of gene size in the precise expression timing of different gene products is not well understood. For optimal cellular efficiency, proteins would be produced only when they were functionally needed. Additionally, when large amounts of genes are being expressed during a cellular response, staggered production of transcripts can ensure that the translational machinery is not overwhelmed. While it has been suggested that gene length acts to establish co-expression of proteins which interact together in the same complex[164], we propose that it also contributes to the coordination of the production of proteins that are involved in the same functional pathways. If gene length plays a role in coordinating expression during the serum response, we would predict that the order in which response genes complete transcription is important. Though gene sizes may change over evolutionary time, the maintenance of relative gene size allows for induced genes to still be expressed in the same order. Additional studies are needed to elucidate the importance of temporal expression of different protein components for optimal signal transduction and the assembly of structural complexes as cells reenter the cell cycle following serum stimulation. However, it is not an easy task to relate transcription timing of response genes to protein functional activity in signaling pathways, as they are not a simple reflection of position within cascades. Additionally, there are several other regulation points that likely influence expression timing, such as RNA processing, translation, post-translational processing, and RNA and protein degradation. Furthermore, many proteins interact with numerous other proteins, so teasing apart the interconnected temporal expression patterns will be difficult.

While gene size influences the amount of time needed to complete transcription, there are still ways in which cells can overcome fixed length restrictions on transcriptional timing. Many genes have multiple transcript isoforms which can be quite different in size. Certain isoforms may be expressed during a cellular response in order to optimize expression timing. Additionally, transcriptional timing may be fine-tuned through modulation of RNAPII elongation rates. We previously concluded that elongation rates are gene specific and correlate with certain histone marks [68]. For some immediately induced response genes, it seems that elongation may occur at a faster rate than under basal conditions. While we cannot rule out that this difference may be related to experimental technique, high elongation rates have been measured following cellular stimulation[67,159]. If faster elongation rates occur following cellular stimulation compared to basal conditions, how elongation rate is regulated under these conditions would need to be explored. Our data also demonstrated that in genes over 200 kb long, elongation rates increased as RNAPII traveled across the gene, which is consistent with previous studies[67,69]. Since elongation rates appear to increase across both upregulated and downregulated genes following serum stimulation, our results suggest that acceleration of transcription elongation is not due to serum activation but rather is an inherent characteristic of transcription elongation. Even so, this does not preclude the possibility that certain genes may demonstrate increased elongation rates as a result of stimulation.

While we speculate that gene length acts as a biological timer for proper expression during the serum response, the consequences of altering gene lengths and therefore the order of gene expression is unknown. The order of gene expression during the serum response may only be important for certain sets of genes or for those involved in specific pathways. The immediately induced genes showed the greatest variance in gene sizes. Because this broad distribution may set up a distinct order of gene expression, we would predict that shuffling expression order would have greater effects on this group of genes. It should be noted that altering the length of a gene in order to specifically modify transcriptional timing is not a menial task. It would be important to maintain regulatory regions and splice site sequences when manipulating intron length. While this type of genetic manipulation has been difficult, especially at the endogenous loci of multiple genes, the development of new technologies for genome engineering, such as the use of the CRISPR-CAS system[165], may allow these questions to be more easily explored.

In conclusion, our study provides a comprehensive profile of the early transcriptional response following serum stimulation. Early serum response genes display a wide range of lengths which influences their transcriptional timing and allows for coordinated expression of functionally related genes. Gene size likely plays an important role in gene expression timing for many different cellular responses.

3.5 Materials and methods

3.5.1 Cell culture and serum stimulation

HF1, hTERT immortalized foreskin-derived human fibroblasts, were grown in MEM supplemented with 10% FBS, L-glutamine, vitamin mix, and antibiotics. Starved cells were grown in the same media minus FBS for 48hrs. For serum stimulation experiments, FBS was added to the media of starved cells (final concentration 10%). Serum addition was followed by immediate 30min bromouridine labeling. Bru-labeling was done for a starved sample and a serum stimulated sample.

3.5.2 Bru-seq

Bru-seq was performed as previously described[105,147]. Briefly, bromouridine (Bru) (Aldrich) was added to the media of starved or serum stimulated cells at a final concentration of 2 mM and incubated at 37 °C for 30 min. Total RNA was isolated by using TRIzol reagent (Invitrogen), and Bru-labeled RNA was isolated by incubation with anti-BrdU antibodies (BD Biosciences) conjugated to magnetic Dynabeads (Invitrogen) under gentle agitation at room temperature for 1h. 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 were carried out as previously described[105,147].

3.5.3 Gene expression and serum response analysis

RPKM (reads per kilobase per million) values were calculated for individual genes over 300bp for starved cells and for the serum stimulated sample. For genes 30kb and under, RPKM was calculated using read counts from the entire gene. For genes over 30kb, an RPKM value was calculated using read counts from the first 30kb downstream of the TSS. Genes were classified as expressed if they had a mean RPKM value greater than 0.5 in starved cells or in the serum stimulated sample.

To identify response genes, an inter-sample comparison analysis was done to obtain RPKM fold change values for each gene in the serum stimulated sample compared to the starved sample[105]. Because we had two replicates of the experiment, the data was combined as described[105]. Serum response genes were defined as genes with a significant change in transcription initiation in stimulated cells compared to starved cells (adjusted p-values <0.05, n=2). P-values were calculated for each gene using DESeq on the combined data. Genes with increased expression were classified as induced and genes with decreased expression were classified as repressed. Repressed genes were required to be expressed (RPKM value > 0.5) in the starved sample.

3.5.4 Ortholog size comparison

Human protein coding genes and corresponding mouse ortholog sizes were obtained using Ensembl Biomart[1]. Genes with one-to-one orthologs were assessed.

3.5.5 Gene enrichment analysis

DAVID was used to analyze functional annotation enrichment. We grouped induced genes into five categories based on length: 0-10 kb, 10-25 kb, 25-50 kb, 50-100 kb, and >100 kb. We grouped repressed genes into two categories: 0-15 kb and >15 kb. For the background gene set for each analysis, we used all expressed genes which fell within the given size range. Here, we report pathways which had a p-value < 0.05.

We also performed gene set enrichment analysis (GSEA) [137]. All genes expressed >0.5 RPKM were rank-ordered according to log2-fold changes compared to the gene expression in serum-starved cells. We utilized the transcription factor targets gene sets.

3.5.6 Elongation analysis

Images displaying read distributions across genes longer than 200 kb were visually assessed to estimate distances travelled by RNAPII in the 30-60, 60-90, and 90-120 minute labeling periods according to the changing positions of the transcriptional wave. 21 genes were analyzed for groups of induced and repressed genes, and median values were recorded for the different labeling periods. These values were used to calculate an average elongation rate for induced and repressed genes during each labeling period.

3.5.7 Statistical analysis

Gene size distributions were analyzed using the Mann-Whitney U test to compare gene sets and the F test to compare variances. P-values < 0.05 were considered significant.

3.5.8 Author information

Kirkconnell, K.S., Magnuson, B., Paulsen, M.T., Lu, B., Bedi, K., Ljungman, M.

KSK, MTP, and BL contributed to experiments. KSK, BM, KB, and ML contributed to data analysis. KSK and ML contributed to figure production and manuscript writing.

CHAPTER IV

Genome-wide mapping of regulatory elements through identification of enhancer RNA transcription using BruUV-seq

4.1 Abstract

Complex interactions of enhancer regulatory elements with genes act to establish the specific expression patterns necessary for cellular identity and diverse responses to external stimuli. The advancement of genome-wide technologies allows for examination of genetic and epigenetic elements, and for cell type specific annotation of putative enhancers. In particular, histone modifications and transcription factor binding profiles have been used to predict the locations of enhancers. However, given the numerous putative enhancer sites identified using these criteria, which outnumber the number of coding genes by an order of magnitude, annotations based solely on these marks may not be the optimal way to classify active enhancer profiles. The widespread detection of transcription occurring at enhancers, especially at super-enhancers, suggests that enhancer RNA may be a robust indicator of enhancer activity.

In this study, we use BruUV-seq to identify genome-wide putative active enhancer elements. BruUV-seq allows us to enrich for enhancer transcripts which are normally very unstable and present at low levels during regular nascent RNA sequencing. We are able to capture enhancer transcripts which align with enhancer marks in untreated cells, and identify changes in enhancer transcription following cellular stimulation. We also examined the response of enhancer activity after inhibition of the transcriptional coactivator and enhancer chromatin regulator BRD4 using the drug JQ1. Treatment of cells with JQ1 has previously been shown to target super-enhancer activity. We see that genome-wide enhancer RNA levels decrease within the first 30 minutes of JQ1 treatment, suggesting that BRD4 maintenance is important at active enhancers. Our data demonstrates the utility of using BruUV-seq to map enhancer RNA and study enhancer activity under basal conditions as well as after stimulation or treatment. In the future, this

technique may be used to identify enhancers or establish cellular identity without using chromatin marks. This study highlights the importance of understanding the biology and function of enhancer transcription, and the need for rigorous investigation of prospective therapeutic agents that target enhancer activity.

4.2 Introduction

Gene expression is controlled by an intricate network of regulatory elements. Enhancers were first identified as genetic elements that contain transcription factor binding motifs and enable increased transcription levels[5,6,166]. The human genome encodes for millions of potential enhancer elements, but only a specific subset of these is active in a particular cell type[6]. Thus, enhancers play a critical role in establishing cellular identity and allow cells to respond to environmental cues. However, the exact mechanistic details behind enhancer selection, activation, and function remain unclear.

Enhancer selection is thought to be established by combinatorial interactions between DNA binding factors. Pioneer transcription factors (TFs) facilitate nucleosome remodeling and allow lineage-determining TFs to bind and prime cell specific enhancers[9,10]. This can lead to subsequent recruitment of collaborative TFs and transcriptional cofactors, such as histone methyltransferases that monomethylate histone H3 lysine 4 (H3K4me1)[27]. These poised enhancers can be activated by further recruitment of TFs, such as signal-dependent TFs[5], and cofactors, such as histone acetyltransferases that acetylate H3K27[167]. Binding of RNA polymerase II (RNAPII) and cofactors such as bromodomain-containing protein 4 (BRD4) can result in the initiation and elongation of transcription at the enhancer[17]. Chromatin looping factors can facilitate interactions between enhancers and promoters[18,168], and it is likely that the collaboration between DNA binding factors at these regions allows for maximum transcriptional output[20].

The characterization of these enhancer features has enabled the genome-wide annotation of putative enhancer elements. Histone marks are commonly used to identify putative enhancers, such as the enrichment of H3K4me1 compared to H3K4me3[11,12], which distinguishes enhancers from promoters, and the presence of H3K27ac[13]. However, annotation using these epigenetic marks still results in exceedingly large numbers of putative enhancers elements, which outnumber the number of coding genes by an order of magnitude[169]. Thus, alternative signatures of enhancers are necessary in order to more accurately annotate active enhancer profiles in cells. Putative enhancers with the classic epigenetic marks are pervasively transcribed into short, non-coding RNAs (eRNAs), and changes in eRNA levels correlate with expression changes of nearby genes[15,16]. Additionally, super-enhancers, which are large enhancer dense regions with high densities of transcriptional coactivators that regulate key drivers of cellular identity[34], have been shown to produce high levels of eRNA[38,39]. This evidence suggests that detection of eRNA transcription may help improve efforts to annotate active enhancer elements.

In addition to being able to identify active enhancers, efforts have been made to modify enhancer activity. For example, super-enhancers drive the expression of oncogenes in cancer cells, and targeting the activity of these super-enhancers can be used to specifically inhibit tumor growth[37]. Super-enhancers contain high densities of BRD4 binding[37], and several drugs have been developed to target BRD4 and other similar chromatin regulators[170]. BRD4 is a bromodomain and extraterminal domain (BET) protein that binds to acetylated histones, plays a role in transcriptional activation, and promotes the elongation of mRNAs and eRNAs[17,171]. An example of a small molecule inhibitor includes JQ1, which binds to the bromodomains of BET proteins with high specificity and prevents binding to acetylated histones[172]. JQ1 has been shown to reduce BRD4 occupancy at super-enhancers and result in decreased target gene expression[37]. Additionally, JQ1 treatment has been shown to inhibit eRNA synthesis[17]. However, it is unclear how drugs such as JQ1, which target important transcriptional cofactors that act globally within the cell, affect genome-wide enhancer activity and transcription.

In this study, we used BruUV-seq to identify active enhancer elements in human cells[109]. Using this technique, which involves the irradiation of cells with UV light prior to Bru-labeling nascent transcripts, we were able to capture and enrich for eRNAs that are normally rapidly degraded. Furthermore, we detected changes in eRNA transcription following cellular stimulation and disruption of enhancer activity using the BRD4 inhibitor JQ1. Our data demonstrate the ability of BruUV-seq to monitor eRNA production and study enhancer activity under different cellular conditions. Annotation of enhancer elements using BruUV-seq can assist future investigations of cell type-specific gene regulation and help to uncover the mechanisms behind the maintenance of enhancer activity.

4.3 Results

4.3.1 Use of BruUV-seq to identify putative active enhancers

We previously observed that irradiation of cells with UVC light prior to metabolic labeling of nascent RNA with bromouridine (Bru) results in redistribution of nascent RNA reads with increased signal at transcription start sites (TSSs) and decreased reads within gene bodies[109]. UV light introduces DNA lesions across the genome that act as transcriptional blocks[110], and so RNA polymerases stall at these lesions within gene bodies but continue to initiate at TSSs[111,112]. Thus, the BruUV-seq technique is useful for genome-wide identification of TSSs. In addition to enrichment of reads at TSSs, we also observed increased signal of antisense promoter upstream transcripts (PROMPTs). PROMPTs are usually unstable and rapidly degraded by the exosome[23,124]. We hypothesize that stalled polymerases may act to protect RNA species that are normally targeted for degradation by shielding the 3'-end of the transcript from the 3'-5' RNase activity of the RNA exosome[173]. UV light has also been shown to inhibit RNA exosome activity[174,175], which could result the enrichment of these transcripts in BruUV-seq data compared to normal Bru-seq data.

Enhancer transcripts are similar to PROMPTs in that they normally display low stability, they are not spliced or polyadenylated, they share similar epigenomic features, and they are targeted by the RNA exosome[6,176,177]. Given these similarities, we expected to be able to detect eRNAs using BruUV-seq. We observed increased signal at a well characterized *FOS* enhancer in the BruUV-seq data compared to the Bru-seq data (Fig. 4.1A[109]). This peak corresponded with characteristic histone marks of enhancers: high H3K4me1 and H3K27ac, and low H3K4me3.   We also identified intergenic regions with high densities of BruUV-seq peaks that coincided with enhancer histone marks and were located upstream of highly expressed genes such as *THBS1* and *MALAT1* (Fig.

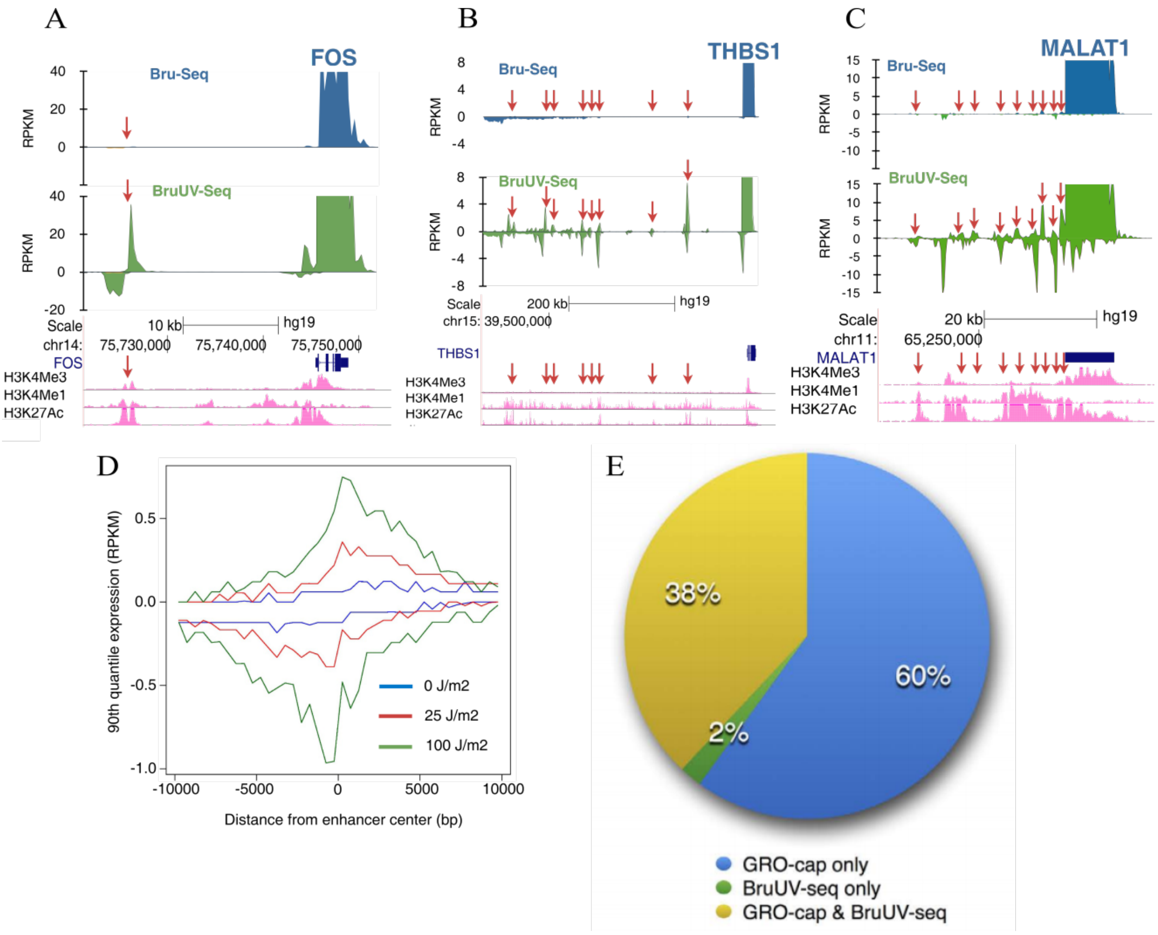


Figure 4. 1: Use of BruUV-seq for detection of enhancer transcription.

(A) Bru-seq (blue) and BruUV-seq (20 J/m^2 UVC) (green) data for the *FOS* gene and upstream enhancer element (red arrow) in HF1 cells. The histone modification tracks are from ENCODE for normal human lung fibroblasts (NHLF). (B-C) Bru-seq and BruUV-seq data for upstream regions of the highly expressed *THBS1* and *MALAT1* genes in HF1 cells, with red arrows indicating potential enhancer peaks. (D) An aggregate view of the reads surrounding 526 intergenic enhancer regions defined by ENCODE genome segmentation annotation in K562 cells. (E) Comparison of intergenic BruUV-seq (100 J/m^2 UVC) and GRO-cap peaks overlapping ENCODE enhancers in K562 cells.

4.1B,C[109]). We propose that these areas may represent active super-enhancers that produce high levels of eRNA transcripts.

We went on to characterize the genome-wide transcriptional signal at enhancers. First, we identified intergenic segments which showed enhanced signal in the BruUV-seq data compared to the Bru-seq data. We then focused on the BruUV-seq identified segments which overlapped with enhancer regions defined by ENCODE[178]. Our genome-wide analysis demonstrated that the signal within annotated enhancer regions is bidirectional and UV dose dependent (Fig. 4.1D[109]). We compared our BruUV-seq data to GRO-cap data[124], a nuclear run-on technique followed by 5'cap sequencing which also detects nascent eRNA transcripts. Comparisons of signal detected within annotated enhancers suggest that Bru-seq identifies a subset of enhancer elements identified by GRO-cap (Fig. 4.1E[109]). Because GRO-cap genomic coverage is more focused, these discrepancies may be due to increased sensitivity of the GRO-cap technique. However, some of these inconsistencies may also be due to methodology as GRO-cap labeling is done *in vitro* as opposed to the *in vivo* labeling done in Bru-seq. Therefore, these GRO-cap identified enhancers may represent the presence of RNAP II at the enhancer and the potential of enhancer activity but not current production of eRNA, though further analysis is needed to explore this possibility.

4.3.2 Gene expression changes are accompanied by changes in eRNA transcription

Because levels of eRNA production often correlate with expression level changes in nearby genes, we sought to use BruUV-seq to observe this relationship between transcription at genes and enhancers. We first used TNF stimulation to induce transcriptional changes through the acute inflammatory response. Using Bru-seq, we were able to identify inflammatory signaling genes which were upregulated after 1hr of TNF treatment, such as *NFKB1, IL8, IL1A,* and *IL1B* (Fig. 4.2A-C[109]). In the BruUV-seq data, we observed peaks upstream of these genes that aligned with the enhancer histone marks and also showed increased signal following TNF treatment. For the *IL8* gene, there were multiple peaks upstream of the gene which all had increased signal following TNF treatment (Fig. 4.2B[109]). The *IL1A* and *IL1B* genes are adjacent to each other, and we observed multiple putative enhancer peaks in the region in between the

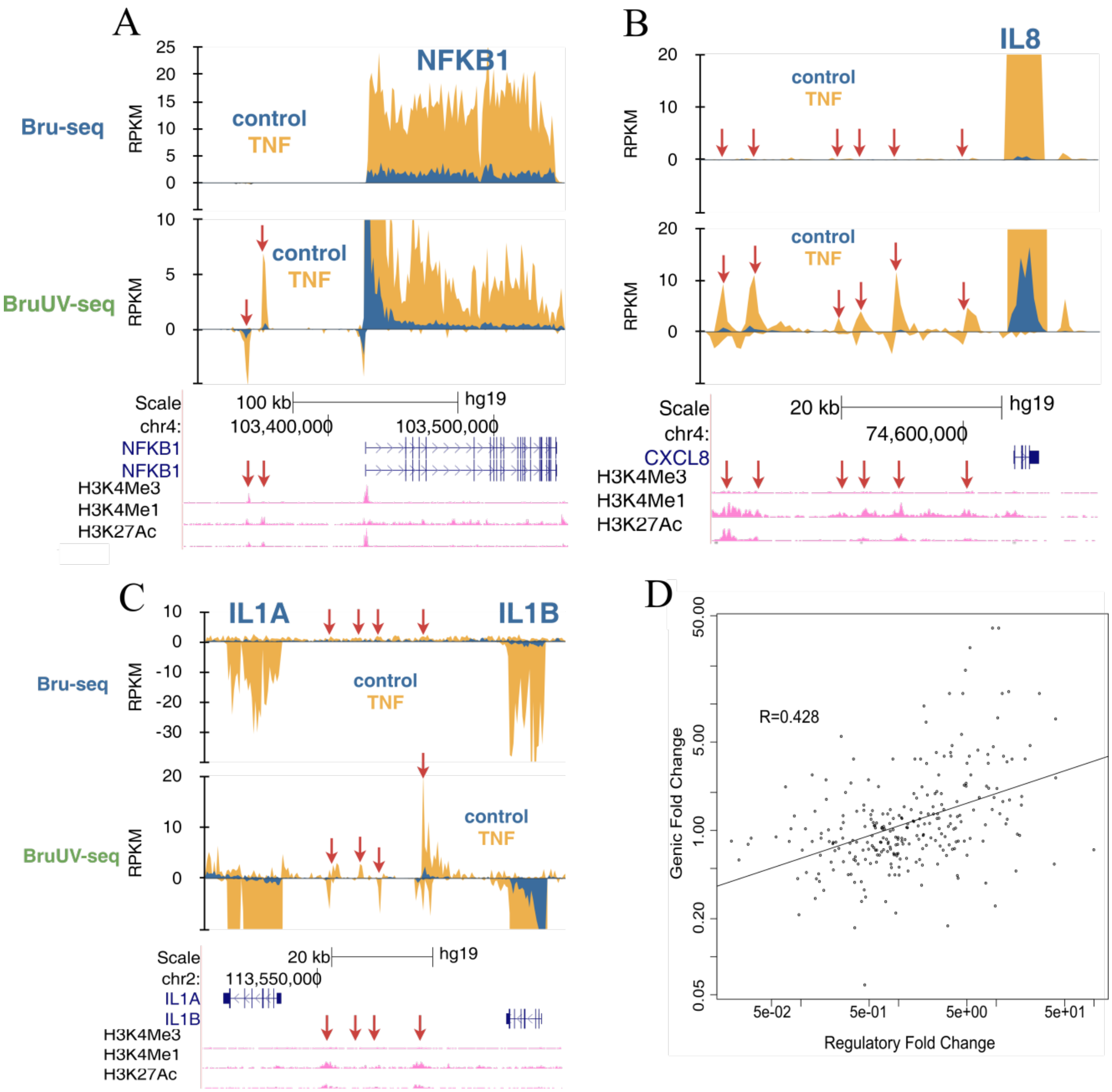


Figure 4. 2: Use of BruUV-seq monitor enhancer activity during TNF stimulation.

Bru-seq (top) and BruUV-seq (20 J/m^2 UVC) (bottom) traces for the (**A**) *NFKB1*, (B) *IL8*, and (C) *IL1A* and *IL1B* genes along with upstream regions in HF1 cells before (blue) and after treatment with TNF for 1 hour (yellow). Enhancer elements activated by TNF are shown with red arrows. (D) Correlation between TNF-induced changes in eRNA and nearest gene mRNA expression in human fibroblasts. The histone modification tracks are from ENCODE for normal human lung fibroblasts (NHLF).

genes (Fig. 4.2C[109]), which suggests that these genes may be regulated by common enhancer elements. Out of the top 99 TNF-induced genes, 65 of these had at least one nearby intergenic BruUV-seq peak, which we interpret to be putative active enhancer elements. Fifteen of these genes had more than five nearby intergenic BruUV-seq peaks. We compared changes in signal at these BruUV-seq peaks to changes in expression levels of the nearest genes and observed a positive correlation (Fig. 4.2D[109]). Because not all enhancers interact with the closest gene, our genome-wide correlation between eRNA activity and gene induction is likely an underestimate.

We also used serum stimulation of starved fibroblasts to induce transcriptional changes. For this experiment, Bru-labeling was done immediately after serum addition, so we were able to observe rapid induction and repression of genes. We identified several upregulated genes which also had corresponding increases in activity at nearby putative enhancer elements. Bru-seq identified a 9-fold upregulation of the *FOS* gene. While in the BruUV-seq data the well-known *FOS* enhancer did not show increased activity as a result of serum stimulation, it appears that an alternative enhancer closer to the TSS may be activated (Fig. 4.3A). Similar findings of increased reads at multiple nearby intergenic peaks were made for the highly induced genes *NR4A1*, *TNC*, *ID1*, *ID2*, and *ID3* (Fig. 4.3B-F). Out of the top 50 most highly induced genes, we found evidence that 39 of these were adjacent to at least one putative activated enhancer element, and out of these, 18 genes were adjacent to 4 or more putative activated enhancer elements (Fig. 4.4). In sharp contrast, only one of the top 15 most highly repressed genes following serum stimulation showed downregulation of nearby putative enhancer elements (Fig. 4.5). Thus, during the serum response, immediate induction of many genes may be associated with rapid activation of nearby enhancer elements, while rapid downregulation is not related to enhancer activity.

4.3.3 Use of bromodomain inhibitor JQ1 to target enhancer activity

Seeing as we were able to use BruUV-seq to identify putative active enhancers, we wanted to validate our method by examining the effects of enhancer activity disruption. We treated HeLa cells with JQ1 for 6hrs prior to performing BruUV-seq.

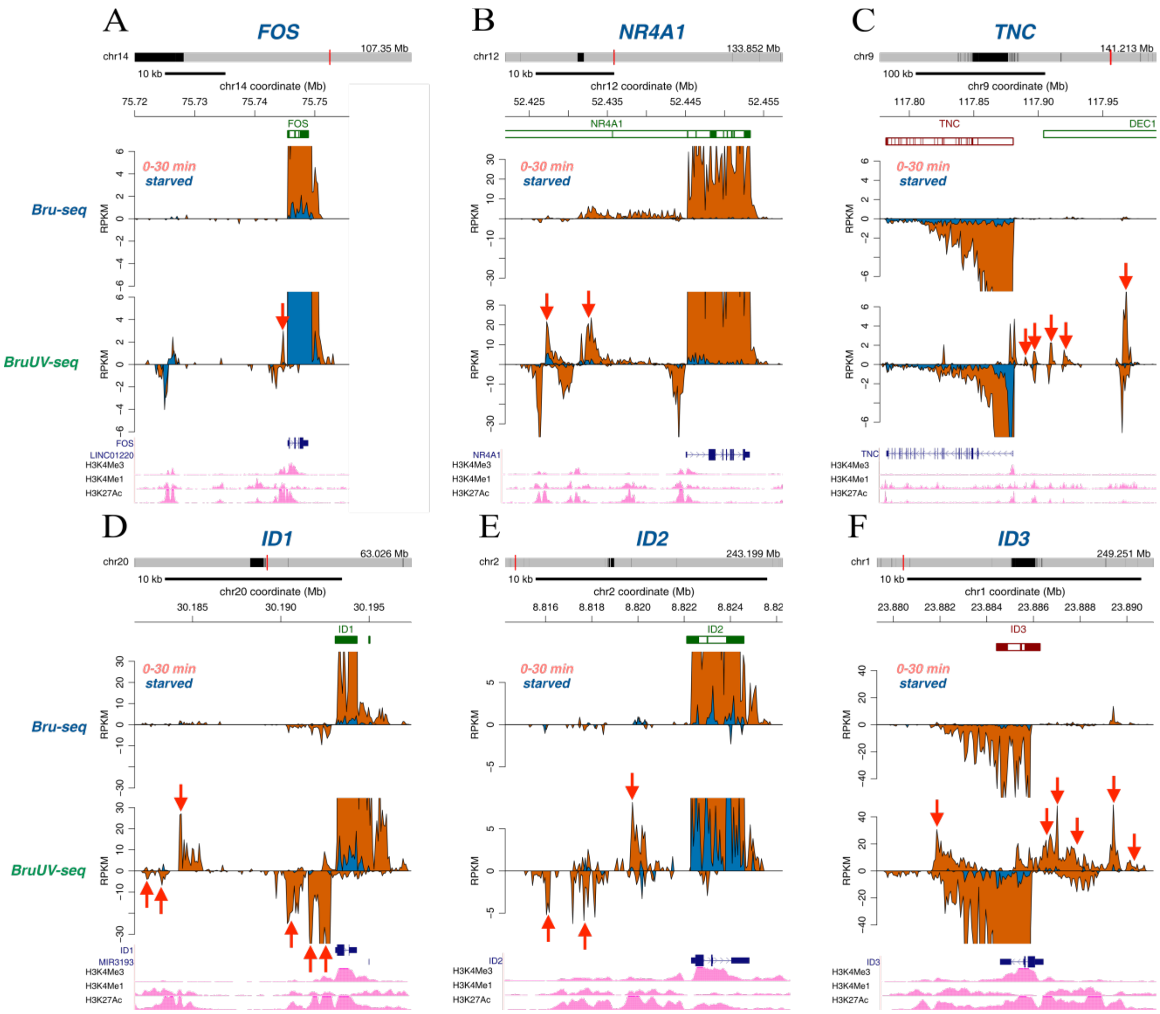
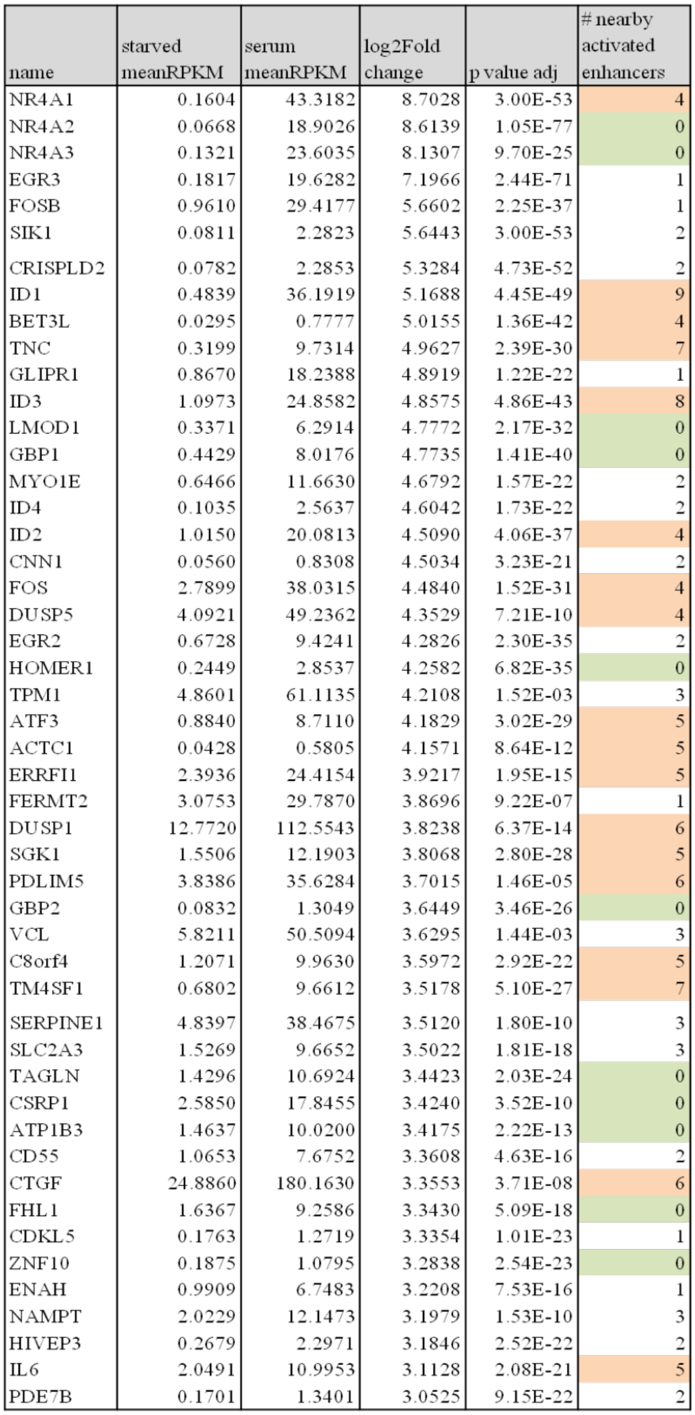


Figure 4. 3: Putative enhancer activation following serum stimulation.

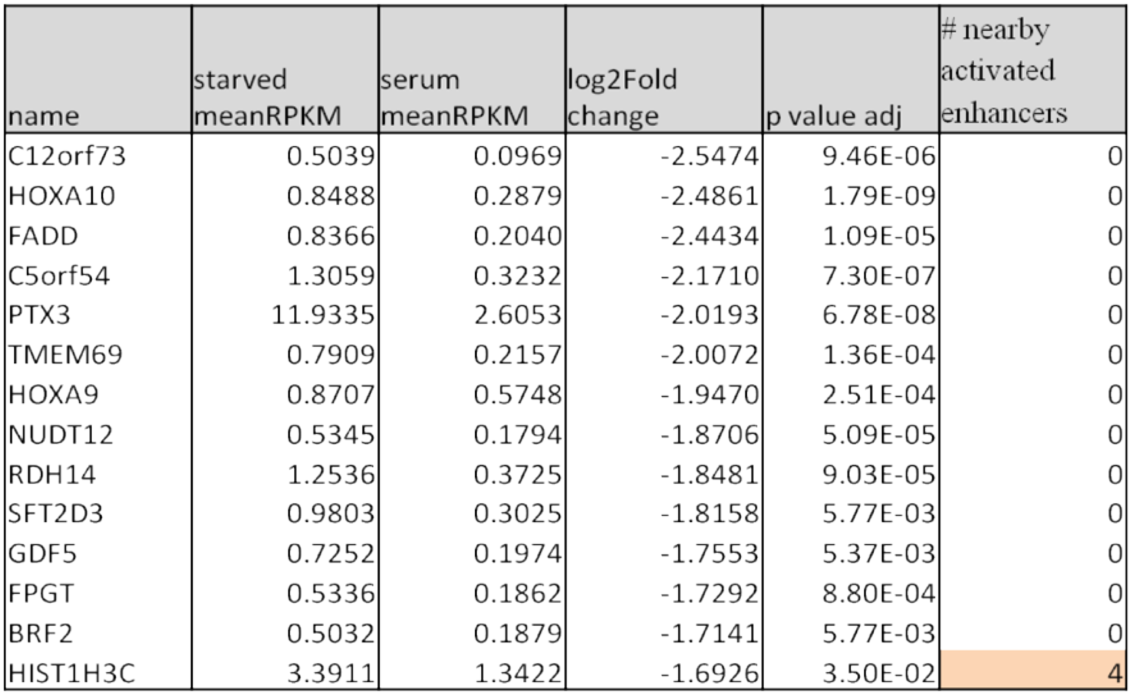
Bru-seq (top) and BruUV-seq (bottom) traces are shown for starved (blue) and serum stimulated cells (orange) for *FOS* (A), *NR4A1* (B), *TNC* (C), *ID1* (D), *ID2* (E), and *ID3* (F). Genes are shown on top in green for plus strand genes and red for minus strand genes. Transcripts and histone mark peaks from ENCODE data for normal human lung fibroblasts (NHLF) are shown below. Red arrows point at enhanced BruUV-seq intergenic peaks that align with peaks for the enhancer marks H3K4me1 and H3K27ac.



name	starved meanRPKM	serum meanRPKM	log2Fold change	p value adj	# nearby activated enhancers
NR4A1	0.1604	43.3182	8.7028	3.00E-53	4
NR4A2	0.0668	18.9026	8.6139	1.05E-77	0
NR4A3	0.1321	23.6035	8.1307	9.70E-25	0
EGR3	0.1817	19.6282	7.1966	2.44E-71	1
FOSB	0.9610	29.4177	5.6602	2.25E-37	1
SIK1	0.0811	2.2823	5.6443	3.00E-53	2
CRISPLD2	0.0782	2.2853	5.3284	4.73E-52	2
ID1	0.4839	36.1919	5.1688	4.45E-49	9
BET3L	0.0295	0.7777	5.0155	1.36E-42	4
TNC	0.3199	9.7314	4.9627	2.39E-30	7
GLIPR1	0.8670	18.2388	4.8919	1.22E-22	1
ID3	1.0973	24.8582	4.8575	4.86E-43	8
LMOD1	0.3371	6.2914	4.7772	2.17E-32	0
GBP1	0.4429	8.0176	4.7735	1.41E-40	0
MYO1E	0.6466	11.6630	4.6792	1.57E-22	2
ID4	0.1035	2.5637	4.6042	1.73E-22	2
ID2	1.0150	20.0813	4.5090	4.06E-37	4
CNN1	0.0560	0.8308	4.5034	3.23E-21	2
FOS	2.7899	38.0315	4.4840	1.52E-31	4
DUSP5	4.0921	49.2362	4.3529	7.21E-10	4
EGR2	0.6728	9.4241	4.2826	2.30E-35	2
HOMER1	0.2449	2.8537	4.2582	6.82E-35	0
TPM1	4.8601	61.1135	4.2108	1.52E-03	3
ATF3	0.8840	8.7110	4.1829	3.02E-29	5
ACTC1	0.0428	0.5805	4.1571	8.64E-12	5
ERRFI1	2.3936	24.4154	3.9217	1.95E-15	5
FERMT2	3.0753	29.7870	3.8696	9.22E-07	1
DUSP1	12.7720	112.5543	3.8238	6.37E-14	6
SGK1	1.5506	12.1903	3.8068	2.80E-28	5
PDLIM5	3.8386	35.6284	3.7015	1.46E-05	6
GBP2	0.0832	1.3049	3.6449	3.46E-26	0
VCL	5.8211	50.5094	3.6295	1.44E-03	3
C8orf4	1.2071	9.9630	3.5972	2.92E-22	5
TM4SF1	0.6802	9.6612	3.5178	5.10E-27	7
SERPINE1	4.8397	38.4675	3.5120	1.80E-10	3
SLC2A3	1.5269	9.6652	3.5022	1.81E-18	3
TAGLN	1.4296	10.6924	3.4423	2.03E-24	0
CSRP1	2.5850	17.8455	3.4240	3.52E-10	0
ATP1B3	1.4637	10.0200	3.4175	2.22E-13	0
CD55	1.0653	7.6752	3.3608	4.63E-16	2
CTGF	24.8860	180.1630	3.3553	3.71E-08	6
FHL1	1.6367	9.2586	3.3430	5.09E-18	0
CDKL5	0.1763	1.2719	3.3354	1.01E-23	1
ZNF10	0.1875	1.0795	3.2838	2.54E-23	0
ENAH	0.9909	6.7483	3.2208	7.53E-16	1
NAMPT	2.0229	12.1473	3.1979	1.53E-10	3
HIVEP3	0.2679	2.2971	3.1846	2.52E-22	2
IL6	2.0491	10.9953	3.1128	2.08E-21	5
PDE7B	0.1701	1.3401	3.0525	9.15E-22	2

Figure 4. 4: Fifty most highly upregulated genes following serum stimulation and nearby putative activated enhancer elements.

Mean RPKM during starved and serum stimulated conditions for two replicates. Adjusted p-values calculated by DESeq. Number of nearby activated enhancers detected using BruUV-seq. Red indicates 4 or more putative enhancers, and green indicates no nearby activated enhancers detected.



name	starved meanRPKM	serum meanRPKM	log2Fold change	p value adj	# nearby activated enhancers
C12orf73	0.5039	0.0969	-2.5474	9.46E-06	0
HOXA10	0.8488	0.2879	-2.4861	1.79E-09	0
FADD	0.8366	0.2040	-2.4434	1.09E-05	0
C5orf54	1.3059	0.3232	-2.1710	7.30E-07	0
PTX3	11.9335	2.6053	-2.0193	6.78E-08	0
TMEM69	0.7909	0.2157	-2.0072	1.36E-04	0
HOXA9	0.8707	0.5748	-1.9470	2.51E-04	0
NUDT12	0.5345	0.1794	-1.8706	5.09E-05	0
RDH14	1.2536	0.3725	-1.8481	9.03E-05	0
SFT2D3	0.9803	0.3025	-1.8158	5.77E-03	0
GDF5	0.7252	0.1974	-1.7553	5.37E-03	0
FPGT	0.5336	0.1862	-1.7292	8.80E-04	0
BRF2	0.5032	0.1879	-1.7141	5.77E-03	0
HIST1H3C	3.3911	1.3422	-1.6926	3.50E-02	4

Figure 4. 5: Fifteen most highly downregulated genes following serum stimulation and nearby putative repressed enhancer elements.

Mean RPKM during starved and serum stimulated conditions for two replicates. Adjusted p-values calculated by DESeq. Number of nearby repressed enhancers detected using BruUV-seq. Red indicates 4 or more putative enhancers detected.

Genome-wide analysis comparing signals within ENCODE annotated enhancer regions in JQ1 treated cells versus untreated cells revealed that JQ1 treatment resulted in an overall decrease in transcription within enhancer regions (Fig. 4.6). While this overall decrease appears moderate, this may be due to JQ1 affecting only a subset of enhancers. We identified 251 genes which were downregulated at least 2-fold following JQ1 treatment. Within the group of downregulated genes, we examined genes with the highest expression during untreated conditions, including *ADAMTS* and *BTBD3* (Fig. 4.7A,B). For these genes we observed multiple BruUV-seq peaks in upstream intergenic regions which also displayed decreased signal after JQ1 treatment. Inhibition of BRD4 with JQ1 appears to have a profound effect on eRNA synthesis at these enhancer elements. In addition to the genes which were downregulated following JQ1 treatment, we also identified small subset of 55 genes which were upregulated at least 2-fold. Within the group of upregulated genes, we examined genes with the highest expression after JQ1 treatment, including *DDIT4 and NEAT1* (Fig. 4.7E,F). While we did observe intergenic BruUV-seq peaks upstream of these genes, the expression of these peaks did not increase following JQ1 treatment. Thus, it seems the downregulation of genes by JQ1 may be through the inhibition of enhancer activity, while the upregulation of genes may be occurring via an alternate mechanism.

We then went on to explore the kinetics of the transcriptional affects of JQ1. Previous studies treated cells with JQ1 for several hours prior to assessing its effects[37,179], however whether this pre-incubation time is necessary in order to observe a decrease in transcription at enhancers and genes has not been previously explored. To examine the immediate effects of JQ1 treatment, we added JQ1 to cells and then immediately performed BruUV-seq. We identified 507 genes which were downregulated at least 2-fold immediately following JQ1 treatment, including genes that were downregulated after the 6 hour treatment such as *ADAMTS* and *BTBD3* (Fig. 4.7A,B), as well as genes that were not downregulated after the 6 hour treatment such as *GPRC5A* and *SLC38A2* (Fig. 5C,D). Interestingly, the upstream intergenic BruUV-seq peaks for all of these genes showed decreased signal during both the 30 min and 6 hr treatments, even if the gene only showed decreased signal during the shorter treatment time (Fig. 4.7A-D). It appears that a large number of genes are immediately downregulated, but then after time gene

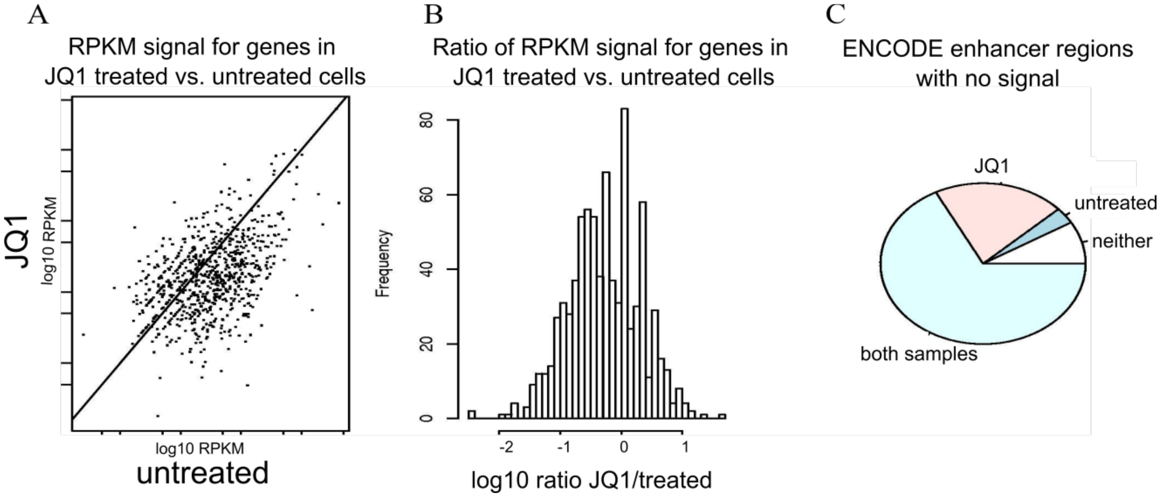


Figure 4. 6: Expression differences at enhancers in untreated and JQ1 treated cells.

HeLa cells treated with DMSO or JQ1 for 6hrs prior to BruUV-seq. (A) Correlation between RPKM signal in untreated and JQ1 treated cells. Each point represents an ENCODE annotated enhancer region which showed expression in both untreated and JQ1 treated cells. (B) Ratio of RPKM signal in JQ1 treated versus untreated cells, with values shifted towards less than 1. (C) ENCODE enhancer regions with no signal in untreated and JQ1 treated cells. More enhancers had no signal in JQ1 treated cells compared to untreated cells.

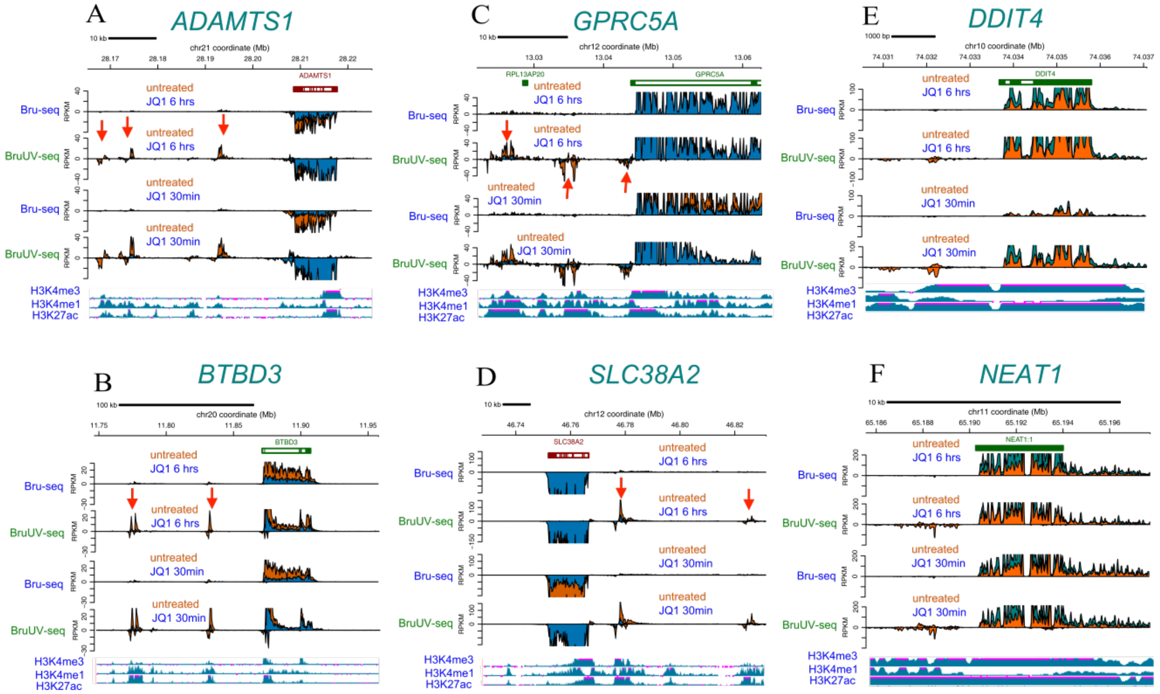


Figure 4. 7: Enhancer transcription following JQ1 treatment.

Bru-seq (first and third rows) and BruUV-seq (second and fourth rows) traces are shown for DMSO (orange) and JQ1 treated (blue) cells for 6hr treatment (first and second row) or 30min treatment during Bru-labeling (third and fourth row). (A-B) Examples of decreased putative enhancer signal and lower expression at the closest gene at both 6hr and 30min timepoints after JQ1treatment. (C-D) Examples of decreased putative enhancer signal at both timepoints, but only lower gene expression at the 30min timepoint. (E-F) Examples of increased gene expression at both timepoints after JQ1treatment, but no change at putative enhancers. Histone mark peaks for HeLa cells from ENCODE data shown below.

expression returns to its original levels for a subset of genes. Because eRNA production remains decreased for these genes, it is possible transcription is able to recover within gene bodies but not at enhancers.

4.4 Discussion

Annotation of enhancer elements based on epigenetic marks alone results in extremely high numbers of putative enhancers, many of which may not be active in a given cell type or under certain growth conditions. Therefore, the ability to distinguish between active and inactive enhancers is important in order to study cell-type specific gene regulation as well as the mechanisms behind enhancer activation. In this study we present the use of the BruUV-seq technique to identify genome-wide nascent eRNA transcription, which is thought to be a defining characteristic of active enhancer elements.

BruUV-seq allows for the detection of normally unstable eRNAs, likely because treatment of cells with UV light results in protection of these transcripts by stalled RNA polymerases and inactivation of the RNA exosome. We compared our data with ENCODE histone modification datasets for our cell lines, and found that the intergenic BruUV-seq peaks aligned with histone mark peaks of typical enhancers. We propose that in the future, BruUV-seq may be used to predict genome-wide regulatory regions without using epigenetic modifications. eRNA transcripts can already be distinguished from mRNA transcripts, even those which are unannotated, based on enrichment during BruUV-seq compared to normal Bru-seq. An important next step will be the ability to distinguish eRNAs from other unstable RNAs, such as long non-coding RNAs (lncRNAs) which are not transcribed at enhancers. This will require a standardized method of how to characterize eRNAs in relation to lncRNAs[6].

To further demonstrate the use of BruUV-seq in identifying active enhancer elements, we examined eRNA transcription following cellular stimulation. Using both TNF and serum to stimulate fibroblasts, we observed intergenic BruUV-seq peaks which increase following these treatments, and that these increases are correlated with the upregulation of nearby gene expression. The detection of changes in eRNA production as a result of cellular stimulation is useful for associating enhancers with their target genes, as we expect to see related increases in enhancer activity and target gene expression. Additionally, we can distinguish active enhancers which are specific to the untreated and stimulated conditions, allowing us to investigate regulation particular to this cellular response. However, because enhancers are not always nearby the gene that they regulate and can be megabase distances away, using the BruUV-seq technique in association with techniques that explore enhancer-promoter interactions and chromosomal looping may be beneficial for assessing genome-wide enhancer-promoter activity.

Lastly, we used BruUV-seq to assess changes in eRNA production after treatment with the BRD4 inhibitor JQ1. We detected decreased enhancer signal at a number of intergenic regions following JQ1 treatment along with the downregulation of nearby genes. This repression of eRNA and gene transcription occurred immediately following JQ1 treatment. While the decreases in enhancer signal were maintained during the 6 hr treatment, a subset of genes appeared to recover from the initial inhibition of gene expression. Our results suggest that maintenance of BRD4 at the enhancer is important for eRNA transcription, and this could be related to the proposed role of BRD4 in transcriptional elongation. However, the relationship between eRNA transcription, enhancer activation, and target gene transcription after JQ1 treatment remains unclear. It is possible that this subset of recovered genes is able to reestablish levels of transcription initiation following disruption of enhancer activity, while initiation of other genes is dependent on enhancer activity. Our data emphasizes the idea that there may be several different classes of eRNA that have different functional roles, and that gene regulation is a complex collaboration between enhancers and other regulatory elements. Additionally, our results stress the importance of understanding how drugs which target chromatin regulators, which play global roles in gene expression, differentially effect subsets of genes.

In conclusion, BruUV-seq is a valuable tool for assessing genome-wide enhancer transcription. Due to its ability to detect changes in eRNA production, this technique will be useful for exploring the functions of eRNA, which are likely to be different depending on the enhancer and target gene. Furthermore, BruUV-seq can be used to identify cell or response specific enhancer elements. Future studies will be critical for untangling the complex networks of gene regulation, and understanding the mechanistic details of eRNA transcription and enhancer activity.

4.5 Materials and methods

4.5.1 Cell culture

HF1 (hTERT immortalized foreskin-derived human fibroblast), HeLa-S3(human cervical adenocarcinoma), and K562 (human chronic myelogenous leukemia) cells were used for the experiments. HF1 cells were stimulated with TNF or serum prior to the Bru-seq and BruUV-seq experiments. For TNF treatment, 10 ng/ml recombinant human TNF-alpha was added to media for 1 hour prior to Bru-labeling or UV irradiation[104]. For serum stimulation, cells were grown in FBS-free media for 48 hours, and then FBS was added back to the media for a final concentration of 10% during Bru-labeling or immediately after UV irradiation. HeLa-S3 were treated with 1uM of the drug JQ1(Cayman Chemical) for 6 hours prior to Bru-labeling or UV irradiation, or during Bru-labeling and immediately after UV irradiation.

4.5.2 Bru-seq and BruUV-seq

Bru-seq was performed as previously described [105,147]. Bromouridine was added to the media of cells at a final concentration of 2 mM and incubated at 37°C for 30 min. Total RNA was isolated using TRIzol reagent, and Bru-labeled RNA was isolated by incubation with anti-BrdU antibodies conjugated to magnetic Dynabeads under gentle agitation at room temperature for 1 h. cDNA libraries were prepared from the isolated 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 were carried out as previously described [105,147].

BruUV-seq was performed as previously described [109]. Cell media was removed and cells were irradiated with a dose of 100 J/m^2 of 254nm UVC light. Immediately following irradiation, cells were Bru-labeled for 30 minutes and the Bru-seq protocol was followed.

4.5.3 Identification of enhancer transcripts

For genome-wide analysis of enhancer expression, BruUV-seq signal was measured within enhancer segments (class E) annotated by the ENCODE genome segmentation for K562 and HeLa-S3 cells[113,178]. BruUV-seq enhancer data was compared to GRO-cap enhancer data[124]. To identify putative active enhancer elements, we identified regions with enhanced signal in BruUV-seq compared to Bru-seq using the previously described UVE analysis[109]. eRNA was distinguished from genic RNA by including only intergenic UVE sites, defined as regions that do not overlap annotated genes or their transcription units. We focused on BruUV-seq peaks that aligned with characteristic enhancer marks using the histone modification tracks generated by ENCODE for the UCSC genome browser for K562 cells or normal human lung fibroblasts (NHLF), the closest cell type match for HF1[113,180].

4.5.4 Author information

Kirkconnell, K.S., Magnuson, B., Veloso, A., Paulsen, M.T., Bedi, K., Ljungman, M.

KSK and MTP contributed to experiments. KSK, AV, BM, KB, and ML contributed to data analysis. KSK, AV, BM, and ML contributed to figure production and manuscript writing:

Figure 4.1: AV, BM, KSK

Figure 4.2: AV, BM, KSK

Figure 4.3: KSK

Figure 4.4: AV, KSK

Figure 4.5: KSK

CHAPTER V

Discussion and future directions

5.1 Transcriptional dynamics and gene expression regulation

Changes in the cellular environment can trigger complex gene expression changes in order to modify cellular function. In Chapter II, I demonstrated the dynamic transcriptional changes that occur during the early serum response. Through examining changes in nascent RNA production, I was able to characterize different transcriptional patterns occurring after serum stimulation. These various patterns of induction and repression illustrate the precise regulation occurring at the transcriptional level. Combined regulation at the stages of transcription, RNA processing, and translation allow for intricate control of gene expression.

5.1.1 Investigating transcriptional dynamics along with other regulatory steps of gene expression

While I presented data on changes in RNA production during the early serum response, this by itself only provides a partial picture of the overall changes in gene expression. Using additional techniques in combination with Bru-seq could be useful for following the complete gene expression process, and to observe how each regulatory step influences later processes.

Bru-seq in conjunction with BruChase-seq can help to analyze the contributions of RNA synthesis and stability during the serum response. Changes in expression levels were not correlated with stability changes during another cellular response[104], and these likely do not have a linear relationship during the serum response either. Furthermore, initial analysis of BruChase-seq data for starved and serum stimulated cells did not indicate a relationship between transcriptional response pattern, either sustained or transient, and stability. This suggests that synthesis and stability is unique to individual genes. By taking both synthesis and stability changes into account, estimations could be made regarding how transcriptional changes would affect the total RNA population. The use of RNA-seq to assess mRNA levels could evaluate the accuracy of the predictions made by Bru-seq and BruChase-seq data. RNA-seq data could also be valuable for examining how transcriptional completion is related to when changes in the total RNA pool can be detected. Increases in mRNA levels may be detected rapidly following the production of full length transcripts for certain genes, while other genes demonstrate additional time delays. This could indicate that post-transcriptional processing steps are playing an important role in expression timing. It would also be important to explore the extent of additional time delays associated with repressed genes, which require degradation of previously made RNAs before transcriptional changes would affect the total RNA pool. Additionally, comparisons of nascent RNA changes to changes in mRNA populations can reveal whether certain changes in transcription initiation have minimal effects on overall RNA levels. Potentially, transient changes in RNA production may not have significant consequences on total mRNA levels or gene expression. It would be important to distinguish how similar patterns in RNA synthesis may affect the expression of genes in different ways.

In addition to combining nascent RNA analysis with techniques that assess RNA at later stages during gene expression, Bru-seq could also be used with methods that assess protein production following serum stimulation. Ribosome profiling, or Ribo-seq, could be used to assess the translatome in addition to the transcriptome[181]. Ribo-seq combines nuclease footprinting with RNA-seq to characterize which RNAs are actively undergoing translation. The BruChase-seq technique could be used along with Ribo-seq to monitor the progression from RNA production to translation over time. Comparisons of RNA expression levels and ribosome occupancy can further distinguish whether translation levels are influenced by RNA abundance or ribosome loading density. Together these techniques could be used to monitor global gene expression from transcription initiation to translation. Similar to the diverse transcriptional dynamics I have described here, post-transcriptional processes will also likely display unique patterns of regulation. To further explore protein production, reverse-phase protein lysate microarrays (RPA) could be used at specific times following serum addition when protein production would be predicted to be completed. RPAs are a microdot western blot that allows simultaneous assessment of selected proteins in multiple samples[182]. All together, future work to explore temporal gene expression dynamics at every stage during a cellular response will provide valuable insights into the multifaceted control of cellular function.

5.1.2 Transcriptional dynamics of individual cells

Bru-seq analysis following serum addition captured the unique transcriptional profile as cells prepared to re-enter the cell cycle from a quiescent state. Interestingly, the transcriptional response appeared to be extremely synchronized, as we could detect distinct waves of transcriptional induction and repression. This suggests that changes in RNA production follow strict temporal patterns. However, whether the changes occur in every cell is unknown. It is possible that out of the hundreds of response genes we identified, an individual cell would only display a subset of these changes. Nascent RNA sequencing within a single cell would be required to robustly analyze how transcriptional changes differ on a cell to cell basis. Comparisons of many individual cells would provide insight into the heterogeneity of the serum response. It could also potentially identify differences in gene expression that result in cells reentering the cell cycle versus those that remain quiescent. While there are currently established protocols for performing single cell RNA-seq, the additional step of isolating labeled nascent RNA produces additional challenges. The amount of nascent RNA at the single cell level is extremely low, less than one picogram. Isolation of these small amounts using antibodies is particularly difficult, and so alternate isolation methods may be necessary. Potentially, sequencing of single cell nuclear RNA could be used as an alternative way to enrich for nascent transcripts. Optimization of single cell protocols to assess nascent transcription will be critical for a deeper understanding of cellular autonomy.

5.2 Gene length and transcriptional timing

The time necessary for producing a full length transcript is inherently linked to gene length. Large variations in mammalian gene sizes result in drastically different time requirements for RNA production, from a few seconds to many hours. In Chapter III, I presented data that illustrates how gene length and their associated transcriptional delays contribute to the regulation of temporal gene expression. Supporting the notion that gene length is important, relative gene length has been maintained in verterbrates[163]. While relative gene length is consistent across species, human genes tend to be larger than other vertebrates. Longer lengths allow for distinct transcriptional timing patterns for many genes that are simultaneously induced.

5.2.1 Manipulating gene length

Further testing of whether gene length is important for setting up precise expression patterns during a cellular response requires perturbation of the normal system. One way to do this would be to change the length of an immediately induced serum response gene so that transcriptional timing is altered. The best way to accomplish this would be to manipulate gene length in a way that does not otherwise perturb gene expression. This makes expression at the endogenous loci preferable. Changes to the endogenous gene can be made using the genomic engineering CRISPR-Cas9 system. This technique uses a guide RNA to target specific genomic sequences and introduce a double strand break (DSB) in the targeted region, which can be repaired via nonhomologous end joining and result in an insertion or deletion mutation. Introduction of two DSBs can be used in order to achieve larger deletions. Alternatively, a repair template can be provided to attain a more precise sequence insertion or deletion. The CRISPR-Cas9 system is a valuable tool that could be used for manipulating the length of individual genes.

A primary experiment would be to make a long immediately induced gene shorter so that completion of transcription occurs more rapidly. Candidate genes for testing may include those with known roles in important serum response pathways that display detectable defects after serum stimulation when they are knocked out. Intronic sequences could be targeted for removal. It would probably be best to avoid complete removal of any introns. Introns can contain regulatory sequences that may influence expression. Additionally, it would be important to maintain RNA processing steps as much as possible, since processes such as splicing are interconnected with gene expression regulation. Any genetic manipulation would need to be tested to ensure that expression levels were not altered. After validating expression levels, it would be appropriate to examine the effects of earlier transcriptional completion of a normally longer gene. Multiple long genes should be shortened and assessed because it is likely that certain genes may show an effect while others do not. Additionally, if gene length plays a role in preventing the translational machinery from being overwhelmed the during the serum response, it may be important to observe the effects of shortening multiple genes at the same time. This may also be accomplished by shortening a transcription factor gene, resulting in earlier activation of target genes. Short genes within the same serum response pathways could also be made longer to increase transcriptional delays. Changing the lengths of multiple genes within the same pathway could indicate whether precise ordering of gene expression is important during the serum response. Overall, genomic engineering provides exciting possibilities for exploring the role of gene length in the regulation of temporal gene expression.

5.2.2 Benefits and risks of gene length

Many increases in gene size are due to intron lengthening[151,162,163]. It is unclear exactly how introns have expanded during evolution, but overall increases in size may be a result of smaller populations in which introns expanded through genetic drift[183]. A large proportion of structural variation originate from the insertion of transposable element (TEs)[184]. Studies have shown that thousands of TEs in the human genome have evolved under strong purifying selection, which suggests that many TEs confer functional advantages[185]. So while TE insertions may not have had immediate adaptive roles, they may have later been co-opted to contribute to cellular function. It has been proposed that TEs play an important role in the evolution of eukaryotic gene regulation[186]. TE insertion can influence gene expression at the transcriptional level in many different ways. New sequences can introduce new promoters, TSSs, or TF binding sites, or they can disrupt existing cis-regulatory elements. Intronic TEs may also drive the production of antisense transcripts that act to attenuate sense transcription. Furthermore, TE insertion can impact gene expression at the level of RNA processing. The addition of sequences can result in alternative splicing, alternative polyadenylation sites, or new miRNA binding sites. Therefore, the lengthening of genes carries the potential for the assembly of new regulatory mechanisms to fine-tune and optimize gene expression levels and contribute to functional complexity.

While increases in gene length may play an important role in diversifying gene regulation, it also comes at a risk. Large genes carry an increased risk for events leading to genomic instability. Many common fragile sites (CFS), which are large chromosomal regions that are sensitive to replication stress and prone to breakage, are within genes larger than a megabase[187,188]. A study in human cells suggested that encounters between the replication and transcription machineries within late replicating long genes results in CFS formation[189]. A more recent study demonstrated that hotspots for copy number variants (CNVs), which are structural alterations including duplications and deletions, occur at the same loci as CFS in a given cell line, and that the active transcription of large genes increases the risk of replication-dependent genomic instability[153]. Therefore, it is important for cells to coordinate efficient transcription and replication to avoid these dangerous chromosomal alterations.

5.2.3 Assessing the coordination of transcription and replication

The coordination of transcription and replication for extremely large genes is not an easy task. Genes over a megabase long can take over 9 hours to transcribe, and so transcription of these genes likely overlaps with S-phase. Under normal cellular conditions, transcription and replication may be regulated so that these processes do not occur at the same time for a given gene. Alternatively, cells may utilize mechanisms which allow them to overcome conflicts between transcription and replication machinery. During head-on encounters between the RNA and DNA polymerases, it is thought that the machineries pause, RNA polymerase is released from the template, and then the replisome progresses through the DNA region[190]. During these events, transcripts may associate with the DNA template and act as a primer for replication. It has been suggested that RNA elongation complexes can be recruited to template-associated transcripts to continue interrupted transcription[189,191]. However, there is not much experimental evidence supporting this model. It may be likely that RNA polymerases must start transcription anew after disruption and release from the template. If this is the case, it would be critical that transcription of large genes have an uninterrupted period in which a full length transcript could be made. During replication, incomplete transcripts of long genes may be generated as a result of disrupted transcription, and then targeted for degradation.

To further explore the coordination of transcription and replication in cells, I attempted to assess nascent RNA and DNA synthesis by developing a Brupli-seq protocol. This method combines the Bru-seq technique with Repli-seq[192], a similar technique that captures newly replicated DNA and can be used to monitor the progression of replication forks over time. By observing transcription and replication profiles during different periods in S-phase, I would see whether these processes occurred within the same genomic regions at the same time. For extremely long genes, this technique would allow me to distinguish whether initiation was temporally restricted, or whether initiation was continuous throughout the cell cycle with the increased risk of producing partial transcripts.

The first step of the Brupli-seq technique was to simultaneously label nascent RNA and DNA. This would be done using Bru for RNA and the thymidine analog 5-ethynyl-2'-deoxyuridine (EdU) for DNA. EdU labeling of DNA allows for selective biotinylation using chick chemistry followed by capture using streptavidin beads. Using EdU as opposed to BrdU allows the cells to be dually labeled due to differing capture methods for RNA and DNA. After labeling, cells would be sorted based on DNA content using flow cytometry. This would allow me to obtain cells in different stages of S-phase. Lastly, RNA and DNA extraction and subsequent isolation of Bru-labeled RNA and EdU-labeled DNA would be performed in order to prepare libraries for sequencing.

Unfortunately, during the development of this technique I encountered several technical challenges. One major challenge involved obtaining cells in different S-phase stages. I wanted to avoid using cell synchronization methods involving chemical blockades to the cell cycle because this results in global perturbations in the cell. Therefore, I opted to use flow cytometry to sort cells which were already in various cell cycle stages. However, because I wanted to isolate RNA after sorting cells, this added some complications. I could not fix cells after staining DNA with DAPI because this would reduce RNA isolation yields. The use of current live cell DNA stains was not an option because the staining protocols required incubation steps that interfered with precise labeling timing. When cells were not fixed, cell morphology was compromised and it was difficult to assess whether cell sorting was resulting in temporally distinct fractions of S-phase cells. While I was able to obtain sorted cellular fractions, additional challenges arose when attempting to isolate Bru-RNA. Bru-seq is typically performed using at least a million cells, but flow sorting for extended periods still resulted in markedly reduced cell counts for each fraction. Attempts to scale down the Bru-seq protocol led to low RNA yields and even less Bru-RNA recovery. I modified the Bru-seq protocol by using a lower RNA-input kit for preparing libraries. However increases in DNA contamination, which were likely related to lower Bru-labeled RNA levels, resulted in data with high genomic background. Therefore, optimization of the Brupli-seq protocol is still required.

One way to circumvent problems associated with cell flow sorting cells is to use an alternative method to get cells at similar cell cycle stages. Development is underway for a microfluidics chip which captures individual cells and allows for new daughter cell to be released and collected. This devise would allow for the collection of synchronized cells without perturbing cellular function or structure. These early G1 cells could then be aged in order to obtain cellular fractions for different S-phase stages. This minimal disturbance to cells could also result in higher RNA recovery levels. Validation of this method should include analysis of the differently aged cell fractions to assess the synchronicity of the cells.

While there is still work to be done on the development of techniques to simultaneously assess replication and transcription, this analysis is important for understanding the coordination of these key cellular processes. Extremely large genes exist in the genome, and how they are efficiently expressed and properly replicated is unclear. While the proposed technique would bring valuable insights into the timing dynamics of RNA and DNA synthesis, it is important to keep in mind that this analysis represents a population of cells. Evidence of concurrent transcription and replication within a genomic region could still be temporally and spatially distinct within an individual cell. For that reason, the ultimate analysis would be done at a single cell level.

5.3 Active enhancer identification

Though the exact functions of eRNA are still not understood, the detection of transcripts produced at enhancer elements is useful for identifying active enhancers and studying enhancer biology. In Chapter IV, I demonstrated the utility of the BruUV-seq technique for capturing eRNA transcription. Further use of BruUV-seq to examine enhancer activity will be important for distinguishing cell type-specific enhancer regulation.

5.3.1 Annotation of active enhancer elements using BruUV-seq

While it is clear that BruUV-seq can detect eRNA transcripts, the development and validation of robust analysis tools are still needed to produce accurate enhancer activity annotations. Our initial method for de novo detection of enhancer transcription involved identification of enhanced BruUV-seq signal within intergenic regions. These regions were defined as those that did not overlap annotated genes. However, this could result in misidentification of unannotated gene TSSs as eRNAs, especially those that represent unstable transcripts. Therefore, the ability to distinguish eRNA transcription units from gene transcription units will be key. Identifying factors may include length and shape of BruUV-seq signal. After designing a model to specifically identify eRNA, we would need to validate a de novo enhancer annotation. Comparison of BruUV-seq identified enhancers to ENCODE defined enhancers, which are based on chromatin signatures, would reveal the accuracy of our model. After we are confident in our model, it could then be used to identify active enhancers in cells which do not have additional information on enhancer signatures.

Enhancer activity profiles will be important for the study of cell type- and cell response-specific gene regulation. BruUV-seq can also potentially be used to identify super-enhancers. Criteria for identifying super enhancers will likely include expression levels and regional peak density. In addition to annotating active enhancers and super-enhancers, it will also be critical to associate active enhancers with gene targets. Genome-wide chromatin conformation capture could be used to provide information on genomic locations that are associated with active enhancers indentified by BruUV-seq, indicating enhancer-promoter interactions. Furthermore, it would be interesting to monitor how chromosome interactions change during a cellular response, and to relate these changes to alterations in enhancer activity. However, it will be important to keep in mind that there are certain limitations to the BruUV-seq technique. BruUV-seq can provide information about eRNA initiation events, but not much additional information about other regulatory transcriptional steps. Additionally, UV treatment induces the DNA damage response in cells, which may affect enhancer transcriptional outputs.

5.3.2 Use of BruUV-seq to analyze potential therapeutics targeting enhancer elements

BruUV-seq is also powerful for monitoring perturbations to enhancer activity. BruUV-seq analysis was able to detect decreases in eRNA transcription following treatment of HeLa cells with the BRD4 inhibitor JQ1, indicating inhibition of enhancer activity. Surprisingly, we observed that these effects of JQ1 occurred immediately after treatment. Concurrent with decreases in enhancer activity was the downregulation of nearby genes. However, it appeared that while enhancer activity remained low, a subset of genes was able to recover expression levels over time. These results indicate that the effects of JQ1 treatment on both enhancers and target gene expression are rapid and complex, and that gene expression changes are not necessarily sustained over time for all genes.

Recognizing and understanding the effects of JQ1 treatment on both cancer and normal cells will be important. JQ1 has been used to treat cancer cells due to preferential targeting of super-enhancer activity and downregulation of key oncogenic drivers[193]. JQ1 and other BET bromodomain inhibitors are currently being explored as potential therapeutic agents, and a handful of these drugs are currently in clinical trials[193]. While these inhibitors appear to selectively downregulate oncogenes, BET inhibitor targets play important roles in genome-wide transcription, and there are likely global affects of these drugs that may be underappreciated. Deeper understanding of the selectivity of genes targeted by BET inhibitors will be critical for predicting overall effects on transcriptional programs. A recent study reported on triple negative breast cancer cell lines that acquired resistance to JQ1 after long term treatment, however it was unclear which mechanisms were responsible for conferring this resistance[194]. Taken together, these results highlight the importance of studying how transcription and cell proliferation occur in a BRD4 dependent, but bromodomain-independent manner.

5.4 Closing remarks

In conclusion, the work presented in this dissertation explored the contribution of early transcriptional events in the regulation of gene expression. Using Bru-seq, I was able to characterize rapid and dynamic transcriptional changes following serum stimulation. Analysis of immediate serum response genes indicated a role for gene length in establishing diverse transcriptional timing patterns. Lastly, using BruUV-seq, I was able to assess genome-wide enhancer activity and identify the immediate effects of BRD4 inhibition on eRNA production and gene expression. This research highlights the power of using the Bru-seq techniques to analyze nascent transcription in order to monitor transcriptional events over time, which will be powerful for future studies on the regulation of cell type- and cellular response-specific gene expression.

REFERENCES

1. Yates, A. *et al.* Ensembl 2016. *Nucleic Acids Research* **44**, D710-D716 (2016).
2. Hatano, A. *et al.* CELLPEDIA: a repository for human cell information for cell studies and differentiation analyses. *Database (Oxford)* **2011**, bar046 (2011).
3. Bulger, M. & Groudine, M. Functional and Mechanistic Diversity of Distal Transcription Enhancers. *Cell* **144**, 327-339 (2011).
4. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotech* **28**, 817-825 (2010).
5. Heinz, S., Romanoski, C.E., Benner, C. & Glass, C.K. The selection and function of cell type-specific enhancers. *Nat Rev Mol Cell Biol* **16**, 144-154 (2015).
6. Li, W., Notani, D. & Rosenfeld, M.G. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat Rev Genet* **17**, 207-223 (2016).
7. He, H.H. *et al.* Nucleosome dynamics define transcriptional enhancers. *Nat Genet* **42**, 343-347 (2010).
8. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).
9. Zaret, K.S. & Carroll, J.S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev* **25**, 2227-41 (2011).
10. Spitz, F. & Furlong, E.E.M. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613-626 (2012).
11. Heintzman, N.D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-318 (2007).
12. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-858 (2009).
13. Creyghton, M.P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences* **107**, 21931-21936 (2010).
14. Mullen, Alan C. *et al.* Master Transcription Factors Determine Cell-Type-Specific Responses to TGF-β Signaling. *Cell* **147**, 565-576 (2011).
15. De Santa, F. *et al.* A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers. *PLoS Biol* **8**, e1000384 (2010).
16. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182-187 (2010).
17. Kanno, T. *et al.* BRD4 assists elongation of both coding and enhancer RNAs by interacting with acetylated histones. *Nat Struct Mol Biol* **21**, 1047-1057 (2014).
18. Kagey, M.H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430-435 (2010).
19. Allen, B.L. & Taatjes, D.J. The Mediator complex: a central integrator of transcription. *Nat Rev Mol Cell Biol* **16**, 155-166 (2015).

20. Hatzis, P. & Talianidis, I. Dynamics of Enhancer-Promoter Communication during Differentiation-Induced Gene Activation. *Molecular Cell* **10**, 1467-1477 (2002).

21. Wang, Q., Carroll, J.S. & Brown, M. Spatial and Temporal Recruitment of Androgen Receptor and Its Coactivators Involves Chromosomal Looping and Polymerase Tracking. *Molecular Cell* **19**, 631-642 (2005).

22. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).

23. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461 (2014).

24. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* **22**, 1775-89 (2012).

25. Ling, J. *et al.* HS2 enhancer function is blocked by a transcriptional terminator inserted between the enhancer and the promoter. *J Biol Chem* **279**, 51704-13 (2004).

26. Ho, Y., Elefant, F., Liebhaber, S.A. & Cooke, N.E. Locus Control Region Transcription Plays an Active Role in Long-Range Gene Activation. *Molecular Cell* **23**, 365-375 (2006).

27. Kaikkonen, Minna U. *et al.* Remodeling of the Enhancer Landscape during Macrophage Activation Is Coupled to Enhancer Transcription. *Molecular Cell* **51**, 310-325 (2013).

28. Lai, F. *et al.* Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* **494**, 497-501 (2013).

29. Li, W. *et al.* Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**, 516-520 (2013).

30. Hsieh, C.-L. *et al.* Enhancer RNAs participate in androgen receptor-driven looping that selectively enhances gene activation. *Proceedings of the National Academy of Sciences* **111**, 7319-7324 (2014).

31. Sigova, A.A. *et al.* Transcription factor trapping by RNA in gene regulatory elements. *Science* **350**, 978-981 (2015).

32. Schaukowitch, K. *et al.* Enhancer RNA Facilitates NELF Release from Immediate Early Genes. *Molecular Cell* **56**, 29-42 (2014).

33. Hnisz, D. *et al.* Super-Enhancers in the Control of Cell Identity and Disease. *Cell* **155**, 934-947 (2013).

34. Pott, S. & Lieb, J.D. What are super-enhancers? *Nat Genet* **47**, 8-12 (2015).

35. Niederriter, A.R., Varshney, A., Parker, S.C.J. & Martin, D.M. Super Enhancers in Cancers, Complex Disease, and Developmental Disorders. *Genes* **6**, 1183-1200 (2015).

36. Whyte, Warren A. *et al.* Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* **153**, 307-319 (2013).

37. Lovén, J. *et al.* Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers. *Cell* **153**, 320-334 (2013).

38. Liu, Z. *et al.* Enhancer Activation Requires trans-Recruitment of a Mega Transcription Factor Complex. *Cell* **159**, 358-373 (2014).

39. Hah, N. *et al.* Inflammation-sensitive super enhancers form domains of coordinately regulated enhancer RNAs. *Proceedings of the National Academy of Sciences* **112**, E297-E302 (2015).
40. Parker, S.C.J. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences* **110**, 17921-17926 (2013).
41. Hnisz, D. *et al.* Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Mol Cell* **58**, 362-70 (2015).
42. Smale, S.T. & Kadonaga, J.T. The RNA polymerase II core promoter. *Annu Rev Biochem* **72**, 449-79 (2003).
43. Lee, T.I. & Young, R.A. Transcription of eukaryotic protein-coding genes. *Annu Rev Genet* **34**, 77-137 (2000).
44. Lee, Y. *et al.* MicroRNA genes are transcribed by RNA polymerase II. *The EMBO Journal* **23**, 4051-4060 (2004).
45. Hernandez, N. Small nuclear RNA genes: a model system to study fundamental mechanisms of transcription. *J Biol Chem* **276**, 26733-6 (2001).
46. Sainsbury, S., Bernecky, C. & Cramer, P. Structural basis of transcription initiation by RNA polymerase II. *Nat Rev Mol Cell Biol* **16**, 129-143 (2015).
47. Hahn, S. Structure and mechanism of the RNA Polymerase II transcription machinery. *Nature structural & molecular biology* **11**, 394-403 (2004).
48. Kim, Y., Geiger, J.H., Hahn, S. & Sigler, P.B. Crystal structure of a yeast TBP/TATA-box complex. *Nature* **365**, 512-20 (1993).
49. Chalkley, G.E. & Verrijzer, C.P. DNA binding site selection by RNA polymerase II TAFs: a TAF(II)250-TAF(II)150 complex recognizes the initiator. *Embo j* **18**, 4835-45 (1999).
50. Oelgeschlager, T., Chiang, C.M. & Roeder, R.G. Topology and reorganization of a human TFIID-promoter complex. *Nature* **382**, 735-8 (1996).
51. Burke, T.W. & Kadonaga, J.T. The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. *Genes Dev* **11**, 3020-31 (1997).
52. Luse, D.S. Promoter clearance by RNA polymerase II. *Biochimica et biophysica acta* **1829**, 63-68 (2013).
53. Westover, K.D., Bushnell, D.A. & Kornberg, R.D. Structural basis of transcription: separation of RNA from DNA by RNA polymerase II. *Science* **303**, 1014-6 (2004).
54. Roeder, R.G. Transcriptional regulation and the role of diverse coactivators in animal cells. *FEBS Letters* **579**, 909-915 (2005).
55. Adelman, K. & Lis, J.T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet* **13**, 720-31 (2012).
56. Jonkers, I. & Lis, J.T. Getting up to speed with transcription elongation by RNA polymerase II. *Nat Rev Mol Cell Biol* **16**, 167-177 (2015).
57. Yamaguchi, Y. *et al.* NELF, a Multisubunit Complex Containing RD, Cooperates with DSIF to Repress RNA Polymerase II Elongation. *Cell* **97**, 41-51 (1999).
58. Wada, T. *et al.* DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes Dev* **12**, 343-56 (1998).

59. Rahl, P.B. *et al.* c-Myc Regulates Transcriptional Pause Release. *Cell* **141**, 432-445 (2010).

60. Barboric, M., Nissen, R.M., Kanazawa, S., Jabrane-Ferrat, N. & Peterlin, B.M. NF-κB Binds P-TEFb to Stimulate Transcriptional Elongation by RNA Polymerase II. *Molecular Cell* **8**, 327-337 (2001).

61. Jang, M.K. *et al.* The Bromodomain Protein Brd4 Is a Positive Regulatory Component of P-TEFb and Stimulates RNA Polymerase II-Dependent Transcription. *Molecular Cell* **19**, 523-534 (2005).

62. Yang, Z. *et al.* Recruitment of P-TEFb for Stimulation of Transcriptional Elongation by the Bromodomain Protein Brd4. *Molecular Cell* **19**, 535-545 (2005).

63. Mueller, D. *et al.* A role for the MLL fusion partner ENL in transcriptional elongation and chromatin modification. *Blood* **110**, 4445-4454 (2007).

64. Peterlin, B.M. & Price, D.H. Controlling the Elongation Phase of Transcription with P-TEFb. *Molecular Cell* **23**, 297-305 (2006).

65. Gilchrist, D.A. *et al.* Pausing of RNA Polymerase II Disrupts DNA-Specified Nucleosome Organization to Enable Precise Gene Regulation. *Cell* **143**, 540-551 (2010).

66. Kwak, H. & Lis, J.T. Control of transcriptional elongation. *Annu Rev Genet* **47**, 483-508 (2013).

67. Danko, Charles G. *et al.* Signaling Pathways Differentially Affect RNA Polymerase II Initiation, Pausing, and Elongation Rate in Cells. *Molecular cell* **50**, 212-222 (2013).

68. Veloso, A. *et al.* Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res* **24**, 896-905 (2014).

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

70. Fuchs, G. *et al.* 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. *Genome Biology* **15**, 1-11 (2014).

71. Saponaro, M. *et al.* RECQL5 Controls Transcript Elongation and Suppresses Genome Instability Associated with Transcription Stress. *Cell* **157**, 1037-1049 (2014).

72. Porrua, O. & Libri, D. Transcription termination and the control of the transcriptome: why, where and how to stop. *Nat Rev Mol Cell Biol* **16**, 190-202 (2015).

73. Skourti-Stathaki, K., Proudfoot, Nicholas J. & Gromak, N. Human Senataxin Resolves RNA/DNA Hybrids Formed at Transcriptional Pause Sites to Promote Xrn2-Dependent Termination. *Molecular Cell* **42**, 794-805 (2011).

74. West, S., Gromak, N. & Proudfoot, N.J. Human 5[prime] [rarr] 3[prime] exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. *Nature* **432**, 522-525 (2004).

75. Zhang, Z. & Gilmour, D.S. Pcf11 Is a Termination Factor in Drosophila that Dismantles the Elongation Complex by Bridging the CTD of RNA Polymerase II to the Nascent Transcript. *Molecular Cell* **21**, 65-74 (2006).

76. Arigo, J.T., Eyler, D.E., Carroll, K.L. & Corden, J.L. Termination of Cryptic Unstable Transcripts Is Directed by Yeast RNA-Binding Proteins Nrd1 and Nab3. *Molecular Cell* **23**, 841-851 (2006).
77. Kim, K.-Y. & Levin, David E. Mpk1 MAPK Association with the Paf1 Complex Blocks Sen1-Mediated Premature Transcription Termination. *Cell* **144**, 745-756 (2011).
78. Ntini, E. *et al.* Polyadenylation site–induced decay of upstream transcripts enforces promoter directionality. *Nat Struct Mol Biol* **20**, 923-928 (2013).
79. Shuman, S. Structure, mechanism, and evolution of the mRNA capping apparatus. *Prog Nucleic Acid Res Mol Biol* **66**, 1-40 (2001).
80. Ghosh, A., Shuman, S. & Lima, C.D. Structural insights to how mammalian capping enzyme reads the CTD code. *Mol Cell* **43**, 299-310 (2011).
81. Hocine, S., Singer, R.H. & Grünwald, D. RNA Processing and Export. *Cold Spring Harbor perspectives in biology* **2**, a000752-a000752 (2010).
82. Wahl, M.C., Will, C.L. & Luhrmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**, 701-18 (2009).
83. Merkhofer, E.C., Hu, P. & Johnson, T.L. Introduction to Cotranscriptional RNA Splicing. *Methods in molecular biology (Clifton, N.J.)* **1126**, 83-96 (2014).
84. Licatalosi, D.D. & Darnell, R.B. RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet* **11**, 75-87 (2010).
85. Chen, M. & Manley, J.L. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10**, 741-754 (2009).
86. Lejeune, F. & Maquat, L.E. Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. *Current Opinion in Cell Biology* **17**, 309-315 (2005).
87. Proudfoot, N. New perspectives on connecting messenger RNA 3' end formation to transcription. *Curr Opin Cell Biol* **16**, 272-8 (2004).
88. Matoulkova, E., Michalova, E., Vojtesek, B. & Hrstka, R. The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells. *RNA Biology* **9**, 563-576 (2012).
89. Kohler, A. & Hurt, E. Exporting RNA from the nucleus to the cytoplasm. *Nat Rev Mol Cell Biol* **8**, 761-773 (2007).
90. Grunwald, D., Singer, R.H. & Rout, M. Nuclear export dynamics of RNA-protein complexes. *Nature* **475**, 333-341 (2011).
91. Wu, X. & Brewer, G. The regulation of mRNA stability in mammalian cells: 2.0. *Gene* **500**, 10-21 (2012).
92. Barreau, C., Paillard, L. & Osborne, H.B. AU-rich elements and associated factors: are there unifying principles? *Nucleic Acids Research* **33**, 7138-7150 (2005).
93. Wilson, G.M. *et al.* Phosphorylation of p40AUF1 Regulates Binding to A + U-rich mRNA-destabilizing Elements and Protein-induced Changes in Ribonucleoprotein Structure. *Journal of Biological Chemistry* **278**, 33039-33048 (2003).
94. Zucconi, B.E. *et al.* Alternatively expressed domains of AU-rich element RNA-binding protein 1 (AUF1) regulate RNA-binding affinity, RNA-induced protein

oligomerization, and the local conformation of bound RNA ligands. *J Biol Chem* **285**, 39127-39 (2010).

95. Jonas, S. & Izaurralde, E. Towards a molecular understanding of microRNA-mediated gene silencing. *Nat Rev Genet* **16**, 421-433 (2015).

96. Chen, C.-Y.A. & Shyu, A.-B. Mechanisms of deadenylation-dependent decay. *Wiley Interdisciplinary Reviews: RNA* **2**, 167-183 (2011).

97. Bracken, C.P. *et al.* Global analysis of the mammalian RNA degradome reveals widespread miRNA-dependent and miRNA-independent endonucleolytic cleavage. *Nucleic Acids Res* **39**, 5658-68 (2011).

98. Parker, R. & Sheth, U. P Bodies and the Control of mRNA Translation and Degradation. *Molecular Cell* **25**, 635-646 (2007).

99. Alwine, J.C., Kemp, D.J. & Stark, G.R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5350-5354 (1977).

100. Weis, J.H., Tan, S.S., Martin, B.K. & Wittwer, C.T. Detection of rare mRNAs via quantitative RT-PCR. *Trends in Genetics* **8**, 263-264 (1992).

101. Hoheisel, J.D. Microarray technology: beyond transcript profiling and genotype analysis. *Nat Rev Genet* **7**, 200-210 (2006).

102. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63 (2009).

103. Ozsolak, F. & Milos, P.M. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* **12**, 87-98 (2011).

104. Paulsen, M.T. *et al.* Coordinated regulation of synthesis and stability of RNA during the acute TNF-induced proinflammatory response. *Proceedings of the National Academy of Sciences* **110**, 2240-2245 (2013).

105. Paulsen, M.T. *et al.* Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA. *Methods* **67**, 45-54 (2014).

106. Tian, B., Nowak, D.E. & Brasier, A.R. A TNF-induced gene expression program under oscillatory NF-κB control. *BMC Genomics* **6**, 1-18 (2005).

107. Anderson, P. Post-transcriptional regulons coordinate the initiation and resolution of inflammation. *Nat Rev Immunol* **10**, 24-35 (2010).

108. Singh, J. & Padgett, R.A. Rates of in situ transcription and splicing in large human genes. *Nat Struct Mol Biol* **16**, 1128-33 (2009).

109. Magnuson, B. *et al.* Identifying transcription start sites and active enhancer elements using BruUV-seq. *Scientific Reports* **5**, 17978 (2015).

110. Sancar, A., Lindsey-Boltz, L.A., Unsal-Kacmaz, K. & Linn, S. Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annu Rev Biochem* **73**, 39-85 (2004).

111. Donahue, B.A., Yin, S., Taylor, J.S., Reines, D. & 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**, 8502-6 (1994).

112. Tornaletti, S. & Hanawalt, P.C. Effect of DNA lesions on transcription elongation. *Biochimie* **81**, 139-46 (1999).

113. ENCODE. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

114. Hon, G.C., Hawkins, R.D. & Ren, B. Predictive chromatin signatures in the mammalian genome. *Hum Mol Genet* **18**, R195-201 (2009).

115. Khodor, Y.L. *et al.* Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes & Development* **25**, 2502-2512 (2011).

116. Caudron-Herger, M., Cook, P.R., Rippe, K. & Papantonis, A. Dissecting the nascent human transcriptome by analysing the RNA content of transcription factories. *Nucleic Acids Research* **43**, e95 (2015).

117. Bhatt, Dev M. *et al.* Transcript Dynamics of Proinflammatory Genes Revealed by Sequence Analysis of Subcellular RNA Fractions. *Cell* **150**, 279-290 (2015).

118. Werner, Michael S. & Ruthenburg, Alexander J. Nuclear Fractionation Reveals Thousands of Chromatin-Tethered Noncoding RNAs Adjacent to Active Genes. *Cell Reports* **12**, 1089-1098 (2015).

119. Churchman, L.S. & Weissman, J.S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**, 368-373 (2011).

120. Mayer, A. *et al.* Native Elongating Transcript Sequencing Reveals Human Transcriptional Activity at Nucleotide Resolution. *Cell* **161**, 541-554 (2015).

121. Core, L.J., Waterfall, J.J. & Lis, J.T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science (New York, N.Y.)* **322**, 1845-1848 (2008).

122. Core, Leighton J. *et al.* Defining the Status of RNA Polymerase at Promoters. *Cell Reports* **2**, 1025-1035 (2012).

123. Kwak, H., Fuda, N.J., Core, L.J. & Lis, J.T. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science (New York, N.Y.)* **339**, 950-953 (2013).

124. Core, L.J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* **46**, 1311-1320 (2014).

125. Lai, F., Gardini, A., Zhang, A. & Shiekhattar, R. Integrator mediates the biogenesis of enhancer RNAs. *Nature* **525**, 399-403 (2015).

126. Tani, H. *et al.* Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Research* **22**, 947-956 (2012).

127. Rabani, M. *et al.* Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nature biotechnology* **29**, 436-442 (2011).

128. Windhager, L. *et al.* Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome Research* **22**, 2031-2042 (2012).

129. Winkles, J.A. Serum- and Polypeptide Growth Factor-Inducible Gene Expression in Mouse Fibroblasts. in *Progress in Nucleic Acid Research and Molecular Biology*, Vol. Volume 58 (ed. Kivie, M.) 41-78 (Academic Press, 1997).

130. Iyer, V.R. *et al.* The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83-7 (1999).

131. Amit, I. *et al.* A module of negative feedback regulators defines growth factor signaling. *Nat Genet* **39**, 503-512 (2007).

132. Tullai, J.W. *et al.* Immediate-early and delayed primary response genes are distinct in function and genomic architecture. *J Biol Chem* **282**, 23981-95 (2007).

133. Herschman, H.R. Primary response genes induced by growth factors and tumor promoters. *Annu Rev Biochem* **60**, 281-319 (1991).

134. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).

135. Huang da, W., Sherman, B.T. & Lempicki, R.A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1-13 (2009).

136. Cooper, L., Johnson, C., Burslem, F. & Martin, P. Wound healing and inflammation genes revealed by array analysis of 'macrophageless' PU.1 null mice. *Genome Biol* **6**, R5 (2005).

137. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).

138. Hess, J., Angel, P. & Schorpp-Kistner, M. AP-1 subunits: quarrel and harmony among siblings. *Journal of Cell Science* **117**, 5965-5973 (2004).

139. Plet, A., Eick, D. & Blanchard, J.M. Elongation and premature termination of transcripts initiated from c-fos and c-myc promoters show dissimilar patterns. *Oncogene* **10**, 319-28 (1995).

140. Liu, X., Kraus, W.L. & Bai, X. Ready, pause, go: regulation of RNA polymerase II pausing and release by cellular signaling pathways. *Trends in Biochemical Sciences* **40**, 516-525.

141. Hasan, N.M., Adams, G.E. & Joiner, M.C. Effect of serum starvation on expression and phosphorylation of PKC-α and p53 in V79 cells: Implications for cell death. *International Journal of Cancer* **80**, 400-405 (1999).

142. Assoian, R.K. & Schwartz, M.A. Coordinate signaling by integrins and receptor tyrosine kinases in the regulation of G1 phase cell-cycle progression. *Current Opinion in Genetics & Development* **11**, 48-53 (2001).

143. Shaulian, E. & Karin, M. AP-1 as a regulator of cell life and death. *Nat Cell Biol* **4**, E131-6 (2002).

144. Hoffmann, E. *et al.* MEK1-dependent delayed expression of Fos-related antigen-1 counteracts c-Fos and p65 NF-kappaB-mediated interleukin-8 transcription in response to cytokines or growth factors. *J Biol Chem* **280**, 9706-18 (2005).

145. Szabowski, A. *et al.* c-Jun and JunB Antagonistically Control Cytokine-Regulated Mesenchymal–Epidermal Interaction in Skin. *Cell* **103**, 745-755 (2000).

146. Shi, Y. *et al.* Starvation-induced activation of ATM/Chk2/p53 signaling sensitizes cancer cells to cisplatin. *BMC Cancer* **12**, 1-10 (2012).

147. Paulsen, M.T. *et al.* Coordinated regulation of synthesis and stability of RNA during the acute TNF-induced proinflammatory response. *Proc Natl Acad Sci U S A* **110**, 2240-5 (2013).

148. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Research* **43**, D662-D669 (2015).

149. Fuchs, G. *et al.* 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. *Genome Biol* **15**, R69 (2014).

150. Saponaro, M. *et al.* RECQL5 controls transcript elongation and suppresses genome instability associated with transcription stress. *Cell* **157**, 1037-49 (2014).
151. Gelfman, S. *et al.* Changes in exon-intron structure during vertebrate evolution affect the splicing pattern of exons. *Genome Res* **22**, 35-50 (2012).
152. Eisenberg, E. & Levanon, E.Y. Human housekeeping genes are compact. *Trends Genet* **19**, 362-5 (2003).
153. Wilson, T.E. *et al.* Large transcription units unify copy number variants and common fragile sites arising under replication stress. *Genome Research* **25**, 189-200 (2015).
154. Swinburne, I.A. & Silver, P.A. Intron delays and transcriptional timing during development. *Dev Cell* **14**, 324-30 (2008).
155. Chorev, M. & Carmel, L. The function of introns. *Front Genet* **3**, 55 (2012).
156. Gubb, D. Intron-delay and the precision of expression of homoeotic gene products in Drosophila. *Developmental Genetics* **7**, 119-131 (1986).
157. Rothe, M., Pehl, M., Taubert, H. & Jackle, H. Loss of gene function through rapid mitotic cycles in the Drosophila embryo. *Nature* **359**, 156-9 (1992).
158. Takashima, Y., Ohtsuka, T., González, A., Miyachi, H. & Kageyama, R. Intronic delay is essential for oscillatory expression in the segmentation clock. *Proceedings of the National Academy of Sciences* **108**, 3300-3305 (2011).
159. Morris, D.P. *et al.* Temporal Dissection of Rate Limiting Transcriptional Events Using Pol II ChIP and RNA Analysis of Adrenergic Stress Gene Activation. *PLoS One* **10**, e0134442 (2015).
160. Fuda, N.J., Ardehali, M.B. & Lis, J.T. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* **461**, 186-192 (2009).
161. Huntzinger, E. & Izaurralde, E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet* **12**, 99-110 (2011).
162. Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B. & Lander, E.S. Human and Mouse Gene Structure: Comparative Analysis and Application to Exon Prediction. *Genome Research* **10**, 950-958 (2000).
163. Yandell, M. *et al.* Large-Scale Trends in the Evolution of Gene Structures within 11 Animal Genomes. *PLoS Comput Biol* **2**, e15 (2006).
164. Chen, X., Shi, S. & He, X. Evidence for gene length as a determinant of gene coexpression in protein complexes. *Genetics* **183**, 751-4, 1si-5si (2009).
165. Sander, J.D. & Joung, J.K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat Biotech* **32**, 347-355 (2014).
166. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299-308 (1981).
167. Stasevich, T.J. *et al.* Regulation of RNA polymerase II activation by histone acetylation in single living cells. *Nature* **516**, 272-275 (2014).
168. Li, W. *et al.* Condensin I and II Complexes License Full Estrogen Receptor α-Dependent Enhancer Activation. *Molecular Cell* **59**, 188-202 (2015).
169. Rivera, Chloe M. & Ren, B. Mapping Human Epigenomes. *Cell* **155**, 39-55 (2013).
170. Chaidos, A., Caputo, V. & Karadimitris, A. Inhibition of bromodomain and extra-terminal proteins (BET) as a potential therapeutic approach in haematological

malignancies: emerging preclinical and clinical evidence. *Therapeutic Advances in Hematology* **6**, 128-141 (2015).

171. Rahman, S. *et al.* The Brd4 Extraterminal Domain Confers Transcription Activation Independent of pTEFb by Recruiting Multiple Proteins, Including NSD3. *Molecular and Cellular Biology* **31**, 2641-2652 (2011).

172. Filippakopoulos, P. *et al.* Selective inhibition of BET bromodomains. *Nature* **468**, 1067-1073 (2010).

173. Lubas, M. *et al.* The human nuclear exosome targeting complex is loaded onto newly synthesized RNA to direct early ribonucleolysis. *Cell Rep* **10**, 178-92 (2015).

174. Blasius, M., Wagner, S.A., Choudhary, C., Bartek, J. & Jackson, S.P. A quantitative 14-3-3 interaction screen connects the nuclear exosome targeting complex to the DNA damage response. *Genes & Development* **28**, 1977-1982 (2014).

175. Tiedje, C. *et al.* p38MAPK/MK2-mediated phosphorylation of RBM7 regulates the human nuclear exosome targeting complex. *RNA* **21**, 262-278 (2015).

176. Scruggs, Benjamin S. *et al.* Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Molecular Cell* **58**, 1101-1112 (2015).

177. Lubas, M. *et al.* The Human Nuclear Exosome Targeting Complex Is Loaded onto Newly Synthesized RNA to Direct Early Ribonucleolysis. *Cell Reports* **10**, 178-192 (2015).

178. Hoffman, M.M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* **41**, 827-41 (2013).

179. Delmore, Jake E. *et al.* BET Bromodomain Inhibition as a Therapeutic Strategy to Target c-Myc. *Cell* **146**, 904-917 (2011).

180. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).

181. Ingolia, N.T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat Rev Genet* **15**, 205-213 (2014).

182. Spurrier, B., Ramalingam, S. & Nishizuka, S. Reverse-phase protein lysate microarrays for cell signaling analysis. *Nat. Protocols* **3**, 1796-1808 (2008).

183. Lynch, M. & Conery, J.S. The origins of genome complexity. *Science* **302**, 1401-4 (2003).

184. Korbel, J.O. *et al.* Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *Science (New York, N.Y.)* **318**, 420-426 (2007).

185. Lowe, C.B., Bejerano, G. & Haussler, D. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proceedings of the National Academy of Sciences* **104**, 8005-8010 (2007).

186. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**, 397-405 (2008).

187. Smith, D.I., Zhu, Y., McAvoy, S. & Kuhn, R. Common fragile sites, extremely large genes, neural development and cancer. *Cancer Lett* **232**, 48-57 (2006).

188. Gao, G. & Smith, D.I. Very large common fragile site genes and their potential role in cancer development. *Cell Mol Life Sci* **71**, 4601-15 (2014).

189. Helmrich, A., Ballarino, M. & Tora, L. Collisions between Replication and Transcription Complexes Cause Common Fragile Site Instability at the Longest Human Genes. *Molecular Cell* **44**, 966-977 (2011).
190. Pomerantz, R.T. & O'Donnell, M. The replisome uses mRNA as a primer after colliding with RNA polymerase. *Nature* **456**, 762-766 (2008).
191. Helmrich, A., Ballarino, M., Nudler, E. & Tora, L. Transcription-replication encounters, consequences and genomic instability. *Nat Struct Mol Biol* **20**, 412-418 (2013).
192. Hansen, R.S. *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc Natl Acad Sci U S A* **107**, 139-44 (2010).
193. Filippakopoulos, P. & Knapp, S. Targeting bromodomains: epigenetic readers of lysine acetylation. *Nat Rev Drug Discov* **13**, 337-356 (2014).
194. Shu, S. *et al.* Response and resistance to BET bromodomain inhibitors in triple-negative breast cancer. *Nature* **529**, 413-417 (2016).