UNIVERSITY OF CALIFORNIA SAN DIEGO

Deciphering principles of regulatory element communication through the analysis of nascent transcription dynamics

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Ryan Patrick Hillary

Committee in charge:

Professor Christopher Benner, Chair
Professor Sven Heinz, Co-Chair
Professor Julie Law
Professor Lorraine Pillus
Professor Francesca Telese

2023

The Dissertation of Ryan Patrick Hillary is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

I dedicate this work to my late father Del Vincent Hillary who passed away in the midst of this research. He was among the very few who expressed true interest in my work. Given our very different scientific backgrounds, he actively worked to understand my project, its motivations and big picture application to the field of biology. He was my motivation and inspiration to pursue hard science, ask hard questions, adhere to quality in research and humbly strive to excel in a niche field. His frequent words of advice and encouragement are deeply missed and carrying on in his absence as well as in the absence of his support has been genuinely difficult.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CRD	Cis-Regulatory Domain
TAD	Topologically Associating Domains
LLPS	Liquid-Liquid Phase Separation
LPS	Lipopolysaccharide
KLA	Kdo2-lipid A
TF	Transcription Factor
TFBM	Transcription Factor Binding Motifs
TLR4	Toll-Like Receptor 4
RE	Regulatory Element
GWAS	Genome Wide Association Study
CRE	Cis-Regulatory Element
TSS	Transcription Start Site
TSR	Transcription Start Region
BMDM	Bone Marrow Derived Macrophages
NGV	Natural Genetic Variation
MPRA	Massively Parallel Reporter Assay
Poly I:C	Polyinosinic-Polycytidylic Acid

ACKNOWLEDGEMENTS

First and foremost, I express genuine thanks to my advisor, Chris Benner, for his patience and counsel as I worked to distill my work here at UCSD into this document. I'm appreciative the time he invested into helping me develop and grow as a scientist. I would also like to thank each of my committee members, Sven Heinz, Julie Law, Lorraine Pillus, Francesca Telese and a previous committee member Olivier Harismendy. These scientists have each provided crucial mentorship during the initial phases and throughout the conclusion of my time here at UCSD.

I'm grateful for the members of the labs of Chris Benner and Sven Heinz, both current and previous for their friendship and help they provided along the way.

Lastly, I would like to thank my family; my wife Katie and my son Clark for putting up with me as I try to balance work and the many difficult life events that occurred during this degree.

Chapters 1,2,4,5,6 represent research intended to be published together pending further insight provided by a Dual-MPRA approach described in Chapter 6. Hillary, R., Guzman, C., Heinz, S., Benner, C. The dissertation author will be the primary investigator and author of this paper.

Chapter 3, as a whole, has been compiled as pending publication. Hillary, R., Heinz, S., Benner, C. The dissertation author will be the primary investigator and author of this paper.

VITA

2014 Bachelor of Science in Bioinformatics, Brigham Young University

2023 Doctor of Philosophy in Bioinformatics and Systems Biology, University of California San Diego

PUBLICATIONS

HM Ollila, E Sharon, L Lin, . . . RP Hillary et al., (2023) Narcolepsy risk loci outline role of T cell autoimmunity and infectious triggers in narcolepsy. *Nature communications* 14 (1), 2709

A Ambati*, R Hillary*, S Leu-Semenescu et al., (2021) Kleine-Levin syndrome is associated with birth difficulties and genetic variants in the *TRANK1* gene loci. *Proceedings of the National Academy of Sciences* 118 (12), e2005753118

OR Phillips, AK Onopa, V Hsu, . . . RP Hillary et al., (2019) Beyond a binary classification of sex: An examination of brain sex differentiation, psychopathology, and genotype. *Journal of the American Academy of Child & Adolescent Psychiatry* 58 (8), 787-798

FE Garrett-Bakelman, M Darshi, SJ Green . . . RP Hillary et al., (2019) The NASA Twins Study: A multidimensional analysis of a year-long human spaceflight. *Science* 364 (6436)

RP Hillary, HM Ollila, L Lin, et al., (2018) Complex HLA association in paraneoplastic cerebellar ataxia with anti-Yo antibodies. *Journal of Neuroimmunology* 315, 28-32

PC Haycock, S Burgess, A Nounu, . . . RP Hillary et al., (2017) Association between telomere length and risk of cancer and non-neoplastic diseases: a Mendelian randomization study. *JAMA oncology* 3 (5), 636-651

*Contributed Equally

ABSTRACT OF THE DISSERTATION

Deciphering principles of regulatory element communication through the analysis of nascent transcription dynamics

By

Ryan Patrick Hilary

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2023

Professor Christopher Benner, Chair
Professor Sven Heinz, Co-Chair

Programs of gene regulation are genetically encoded by regulatory elements, which are short stretches of DNA that typically harbor one or more binding sites recognized by transcription factors. During development or cellular stimulation, activated transcription factors bind to these regulatory elements where they subsequently recruit additional factors to initiate or modulate transcription. In most cases, the precise control of gene expression is dependent on multiple regulatory elements located in the vicinity of a gene, where the prevailing theory describes transcription as being initiated through looping of DNA, physically connecting a distal regulatory element (enhancer) to its intended proximal regulatory element (promoter)

thereby mediating transcription regulation. However, methods that measure nascent transcription have revealed a substantial amount of transcription activity occurs at promoter-distal regulatory elements, often far from the genes themselves. It remains unknown how these many individual distal regulatory elements communicate to drive changes in gene expression.

My thesis work reveals adjacent regulatory elements are induced and repressed simultaneously upon cellular activation, forming cis regulatory domains (CRDs), spanning up to hundreds of kilobases. This observation is at odds with the current promoter-enhancer looping theory, which does not account for the behavior of *all* distal regulatory elements near regulated loci but only specific enhancer/promoter interactions. My thesis data suggest that communication between elements may not depend on specific pairwise regulatory element (RE) interactions but rather are mediated by mechanisms that span across the whole domain, such as enrichment in local TF concentrations or general subnuclear localization observed within CRDs. Our results identify that the loss of 3D genome structure causes a loss of synergistic transcriptional bursts within cis-regulatory domains (CRDs), demonstrating the importance of maintaining an intact genomic regulatory landscape. We report that genetic variation at TLR4 associated transcription factor binding motifs can disrupt the transcriptional landscape within CRDs at both TSS harboring the mutation as well as adjacent TSS within 50 kb of mutated sites. These results represent crucial evidence of communication between regulatory elements and identification of CRDs as genuine genomic entities. Our data supports a data mode of gene regulation that could impact how we study mechanisms governing transcription initiation and modulation of gene expression.

Chapter 1 Introduction

Distal regulatory elements, often referred to as enhancers, are responsible for the recruitment of transcription factors (TFs) that bind to the regulatory sequences they contain (Karnuta & Scacheri, 2018). These transcription factors bound to regulatory elements are believed to be the direct modulators of expression of associated nearby genes (Levine, 2010, Karnuta & Scacheri, 2018). It is still unknown how and if these regulatory elements (REs) individually communicate with one another to drive changes in the transcription of target genes (Field & Adelman, 2020). Previous work has shown that proximal and distal regulatory elements are both transcribed and share similar sequence composition and core promoter elements (Andersson et al., 2014, Kim et al., 2010, Core et al., 2014, Mikhaylichenko et al., 2018). Currently, it is assumed that the activation of enhancers precedes and is required for direct transcription initiation of proximal regulatory elements at the 5' end of genes (promoters) leading to their expression (Blackwood, 1998, Bulger & Groudine, 1999). The mechanisms by which multiple distal regulatory elements communicate to effectuate the dynamic temporal modulation of transcription at these gene bodies continues to be elusive (Ryan & Farley, 2019). This suggests that sequence characteristics of these regulatory elements are the attributes driving their ability to recruit and regulate gene expression (Jindal & Farley, 2021). Non-coding regions of the genome are proving to be crucial in understanding gene expression modulation in the context of disease mechanisms in genome-wide association studies (Rickels & Shilatifard, 2018). While many GWAS risk loci have been identified and the majority of risk sites reside in non-coding regions of the genome, the underlying mechanisms causing disease states remain elusive (Field & Adelman, 2020). These non-coding regions, that contain enhancers, are becoming more prevalent in understanding and controlling gene expression in potential therapeutics (Thakore et al., 2015). Together, these many

studies highlight a need to more fully characterize mechanisms regulating enhancer activity in relation to gene expression and modulation.

One of the earlier models proposed that attempts to characterize enhancer/promoter mediated gene expression and modulation is the looping theory (Blackwood & Kadonaga, 1998, Bulger & Groudine, 1999). This theory describes gene expression as being the direct result of enhancers, after recruiting TFs, looping to physically bind a promoter to mediate the recruitment of additional transcription machinery, including RNA polymerase II to transcribe the gene body itself (Levine, 2010). Recent studies have revealed problems with current models of transcription regulation (Li et al., 2020, Alexander et al., 2019, Delaneau et al., 2019). One notable study examined gene expression after the loss of cohesin, a protein complex that mediates 3D looping. 3D loops in chromatin are thought to be key in facilitating direct enhancer/promoter interactions (Li et al., 2020). After the removal of cohesin, and the observed loss of 3D structure, normal gene expression was observed. One other study using single-gene imaging during transcription initiation revealed that the distance between enhancers and target promoters decreases with activation, however, the distance remaining between them during active transcription does not show these elements coming in direct contact with each other but rather formed a nano-environment rich with multiple regulatory factors upon transcription activation (Li et al., 2020). Another study examined that with respect to the *SOX2* gene, no changes in enhancer-promoter distances were observed during transcription initiation and that regulation of the gene was not dependent on 3D changes (Alexander et al., 2019). Of particular interest was the observation of a genetic phenomenon that was detected when comparing epigenetic dynamics within many individuals of different genetic backgrounds (Delaneau et al., 2019). This genetic phenomenon was characterized as multiple active and adjacent regulatory elements that are all co-regulated *en masse*, such that the functional

role of each individual element is difficult to discern. These studies highlight intriguing observations that challenged our current understanding of the enhancer/promoter looping theory and highlight that we need to more carefully characterize transcription induction and regulation genome wide.

New methods sequencing nascent RNAs that capture both stable and unstable transcripts generated from transcription at both enhancer and promoter sites provide a promising means of effectively deconvoluting transcription mechanisms and potential regulatory element communication driving gene expression (Duttke et al., 2019). The use of capped-short RNA-seq (csRNA-seq), an improved method of capturing and sequencing initiating nascent RNAs with high temporal and spatial resolution, has yielded preliminary results in our studies that continue to challenge the current promoter-enhancer looping theory. The extent to which we detect regulatory elements that are located adjacent to one another and simultaneously activated would make direct contact modulation of promoter site by all distal (enhancer) sites unlikely. This detection of correlated transcription activity demonstrates that regulation is shared across all elements within local domains, suggesting mechanisms regulating gene expression other than just single promoter-enhancer looping. Such an observation within our data, along with other studies calling the current enhancer-promoter looping theory into question led me to explore alternative mechanisms governing transcription initiation and modulation.

An alternative model we propose based on our experimental observations is that specific REs act as the direct means of recruiting and maintaining an optimal concentration of transcription machinery within their genomic loci. As a result, adjacent REs along with gene bodies are also transcribed within the TF rich environment created. This idea moves away from the theory that promoters must come in physical contact with TF bound enhancers and instead suggests that

multiple distal regulatory elements collectively integrate regulatory signals and/or communicate to maintain and modulate transcription regulation within a TF rich environment. This hypothesis is consistent with the recently proposed liquid-liquid phase separation model, where proteins, transcription factors and other transcription cofactors and regulators form genomic compartments of regulation (Sabari et al., 2018). If we can successfully characterize this alternative model, it will better guide our understanding of these non-coding regions and provide the means to further unravel how gene regulation is governed by distal regulatory elements.

Acknowledgements

Chapters 1,2,4,5,6 represent research intended to be published together pending further insight provided by a Dual-MPRA approach described in Chapter 6. Hillary, R., Guzman, C., Heinz, S., Benner, C. The dissertation author will be the primary investigator and author of this paper.

Chapter 2 Leveraging csRNA-seq to Capture the Temporal Dynamics of Robust Immune Responses

2.1 Abstract

The study of genome function within biological systems typically employs methods to capture products and byproducts of genomic activities, one key focus being transcription. Two key genomic entities that transcription is known to be controlled by are enhancers and promoters. To study how enhancer and promoters work, we can employ methods that directly measure their transcriptional activity with high sensitivity which might allow us to learn more about how they communicate with one another. Here, we report a time series approach employing the recently developed csRNA-seq to capture the robust genomic activation of the immune response pathways exhibited within LPS-stimulated macrophage cells. The use of csRNA-seq permits precise measurements of the frequency and location of transcription initiation of both stable and unstable transcripts genome-wide. Because csRNA-seq only requires total RNA that can be rapidly isolated, it makes measuring transcription within tight time intervals possible, overcoming the limitations of other common transcription profiling techniques such as ChIP-seq or GRO-seq. The use of csRNA-seq within a tight time series identified the precise moments of specific transcription factor action that aids in deconvoluting transcription factor timing and duration of activity within the genomic regulatory cascades of the TLR4 pathway. In conclusion, we find this profiling technique to be an effective means of capturing, examining and deconvoluting in-vivo transcriptional mechanisms within an active and complex biological system.

2.2 Introduction

Powerful genomic profiling approaches have recently been applied to decipher regulatory control mechanisms driving transcriptional initiation and activity at individual regulatory

elements. However, these methods either only measure stable transcripts usually initiated at promoters (i.e., CAGE, RNA-seq) (Kodzius et al., 2006, Mortazavi et al., 2008) or profile properties of chromatin (e.g., ATAC-seq, ChIP-seq) (Buenrostro, Wu, Chang, & Greenleaf, 2015). As such, they have been limited by the resolution they can accurately measure transcription activity at clearly defined regulatory elements. For our approach we will utilize csRNA-seq (Duttke et al., 2019). By sequencing the capped short RNAs in a sample (similar to Start-seq (Nechaev et al., 2010)) we can identify nascent initiation at promoters and enhancers. csRNA-seq can accurately capture transcription occurring at specific time points due to the fact that total RNA from samples can be immediately captured without the need of cellular manipulation such as nuclei isolation, providing a high temporal resolution conducive of short interval time course approaches (Duttke et al., 2019).

By profiling as many transcription initiation sites for both stable and unstable RNAs as possible using csRNA-seq we obtain a comprehensive view of transcriptional activity occurring at *all* regulatory elements in the genome, including promoter-distal enhancer regions. Furthermore, csRNA-seq combines these attributes with the ability to report transcription occurring a spatial resolution of 1 nucleotide, precisely identifying transcription start sites (TSS). The use of csRNA-seq supersedes the common use of ATAC-seq/open chromatin methodologies along with histone modification mapping by ChIP-seq in that it is a more direct measurement of transcription, while the aforementioned methods measure other aspects of chromatin that are only correlated with transcription. Also of note is the high dynamic range in which transcription activity is captured using csRNA-seq, ensuring confident assessments of activation/repression.

By leveraging these csRNA-seq characteristics within a tight time interval time series approach we defined precise windows of transcription initiation and regulation at regulatory elements hosting specific transcription factor binding motifs, genome-wide.

2.3 Methods

2.3.1 Experimental System – Toll-Like Receptor 4 Signaling Pathway

Toll-Like Receptor 4 (TLR4) signaling in macrophages is the ideal system to study transcription regulation due to the robust, high-magnitude transcriptional responses that are elicited after activation. (**Figure 2.1**) Once activated, macrophage cells exhibit an acute physiological change to a proinflammatory phenotype that includes the induction of cytokines, chemokines, and other mediators of innate immunity (Fang et al., 2017). The specific use of murine RAW264.7 macrophages was selected due to their ease of transfection and high sensitivity to stimuli (Fang et al., 2017). Previous studies have demonstrated that an acute immune response within macrophages is detectable within one hour of stimulation with endotoxin Kdo2-lipid A (KLA), a purified component of lipopolysaccharide (LPS) recognized by the TLR4 receptor (Tong et al., 2016). Within this immune signaling pathway we are interested in where transcription factors first begin interacting with the genome, the TLR4 associated transcription factor binding motifs are NF-κB, AP-1 and IRF like motifs. This robust response of KLA-stimulated macrophages enables a short-interval time course approach where we can capture the entirety of the transcriptional dynamics of the well-characterized TLR4 pathway on a genomic scale.

2.3.2 Experimental Design – csRNA-seq Short Interval Time Series

Capturing and sequencing nascent RNAs represents the most effective means of accurately measuring transcription dynamics. For the task of identifying individual regulatory elements and their transcriptional activity, we used capped short RNA-seq (csRNA-seq), an assay that has

similar properties to Start-seq and true nascent RNA assays such as GRO-cap/5'GRO-seq (Nechaev et al., 2010, Core et al., 2014). csRNA-seq sequences the 5' end of nascent RNAs produced by RNA polymerase II during transcription initiation. (**Figure 2.2**) As RNA Polymerase II is recruited and binds specific sites of DNA it proceeds to transcribe DNA for roughly 20-60 bp before pausing. The resulting capped small RNAs are produced and readily detected at the TSS of both stable (e.g., mRNA) and unstable (e.g., enhancer RNAs or eRNAs) transcripts, and are proportional to the rates of initiation as detected by nascent transcription assays (Heinz et al., 2010, Nechaev et al., 2010, Core et al., 2014). csRNA-seq is crucial to our study in that these captured nascent RNAs identify the precise genomic location of all individual regulatory elements that are transcribed. For our model dataset, we generated a csRNA-seq time course with KLA-stimulated mouse macrophage cells at 1 min, 5 min, 10 min, 15 min, 20 min, 30 min, 40 min, 50 min, 1 hr, 2 hr, 3 hr, 4 hr, 6 hr and 8 hr after KLA treatment. (**Figure 2.3**) We complemented these experiments with ChIP-seq for PolII, H3K4me2, and H3K27ac, to detect changes in chromatin modifications and RNA polymerase II binding after KLA stimulation. This data was collected as a means to provide an initial characterization of the TLR4 signaling response pathway.

2.3.3 A Versatile Experimental Readout: The Transcription Activity Profile

An assessment of differential expression within our dataset was necessary to ensure expected TLR4-associated genes are differentially regulated and associated motifs are enriched within our model system. The initial quality control and preliminary analyses for each of the sequencing methods was performed using HOMER following established workflows for both ChIP-seq and csRNA-seq (Heinz et al., 2010; Duttke et al., 2019). The use of HOMER streamlined the preliminary steps such as adapter trimming, alignment and peak calling. Sequence alignment was completed using BOWTIE (Langmead, Trapnell, Pop, & Salzberg, 2009) or STAR (Dobin et

al., 2012) with mm10 as the reference genome and findPeaks (HOMER) (Heinz et al., 2010) was used to identify individual peaks, representing strand specific transcription occurring at single-nucleotide resolution, within each timepoint sample (**Figure2.4**). We used ChIP-seq to verify the csRNA-seq results and investigate how chromatin and RNA polymerase II dynamics change at individual sites after KLA stimulation. (Figure 2.3) The main focus of this analysis was to leverage csRNA-seq to temporally characterize the induction patterns of transcriptional activity at specific regulatory elements upon KLA stimulation. To do this we performed a differential expression analysis using DESeq2 (Love, Huber, & Anders, 2014). comparing each timepoint against untreated cells. We detected transcription activity dynamics at a 2-fold change in magnitude for either enrichment or depletion at 12,838 promoter/proximal sites and 57,089 distal/enhancer sites.

2.4 Results

2.4.1 Differential Transcription Activity: Precise Windows of Regulatory Activity

We identified groups of TSSs whose transcription activity dynamics were similar within our time course experiment. By so doing we will obtained an unbiased view of the temporal dynamics of activity at regulatory elements and collect initial evidence of potential coregulation. To collect the continuous transcriptional activity occurring at individual regulatory elements we used HOMER (annotatePeaks) (Heinz et al., 2010) to collect all expression measurements at each timepoint for regions detected to be three-fold enriched or depleted within at least one timepoint (69927 total sites). We clustered the normalized transcription activity measurements at each site using rlog regularization (Love, 2014) and used hierarchical clustering to aggregate sites with similar activity patterns across time points. (**Figure 2.5**) To account for the differences in magnitude and group transcription activity based on induction/repression patterns alone, all individual activity patterns were z-score-normalized and individual clusters were formed based on

Euclidian distance and Ward's method was the criterion applied during clustering (Gower, 1985, Ward, 1963). The results of the hierarchical clustering analysis of the time course data revealed 11 distinct activity patterns that capture unique temporal dynamics of regulatory element activation and deactivation in KLA-stimulated macrophages. (**Figure 2.5**) Within the temporal dynamics observed we report a depression of transcription activity in clusters 4,5,8 and 9. Notably, within the groups 1,2,3,6,10 that show an induction of transcriptional activity upon KLA stimulation, we detect clusters of distinct timeframes of transcription induction during the activation of the TLR4 pathway (Clusters 1 at 2hr, 6 at 40min, 10 at 30min). (**Figure 2.5**)

TLR4 responses should reveal the enrichment of inflammation-associated transcription binding motifs at activated regulatory elements (Zhao et al., 2014, Fujioka et al., 2004). We began to explore the sequence features contained within our grouped clusters of CREs with similar transcription activity using HOMER (Heinz et al., 2010). By so doing we detected strong enrichment for the motifs recognized by NF-κB (p=1e-299, Cluster 10 [n=14448]), AP-1 (p=1e-586, Cluster 6 [n=10208]) and IRF3 (p=1e-75, Cluster 1 [n=5963]) within specific time frames during the progression of the regulatory response. (**Figure 2.6**) Within the TLR4 pathway it is known that the activation of NF-κB and AP-1 occurs first followed by the activation of IRF family transcription factors later in the response (Cheng et al., 2017). The results of our preliminary analysis replicate this general timing. However, due to the high temporal resolution of our data, we can detect a slight but distinct difference in timing between the enrichment of transcription of sites containing NF-κB (30mins), AP-1 (40-50mins) and IRF (1.5-2hr), suggesting a short time interval shift from one TF driving transcription modulation to another, genome wide. To examine the biological aspect of the TSS involved with clusters identified we performed gene ontology (GO) term enrichment analysis. (**Figure 2.6**) The results for cluster 10 were enriched for activity

near genes responsible for defense against bacterial pathogens. Cluster 1 GO analysis results were in line with expected genes regulating immune responses. The enrichment of TLR4 associated within our induced TSS represents signifies a successful induction of this immune pathway such that we can leverage these specific TSS to further investigate genomic attributes driving localized transcription initiation.

2.4.2 Characterization of Localized Transcription Activity – TNFAIP3 Gene

We next wanted to address if there was a relationship between gene proximal and distal regulatory elements due to adjacency in the genome. If similar transcription activity profiles were concentrated locally there could be functional interactions could be implicated. We first examined a known TLR4-associated gene, TNFAIP3, and any distal regulatory elements surrounding this gene body with csRNA-seq signal at any time point to examine the local transcription dynamics. To compare activity patterns at sites of transcription we determined their similarity by Pearson correlation coefficient. . This correlation analysis consisted of comparing each regulatory element's transcription activity pattern to the patterns of all other regulatory elements genome wide. This correlation analysis revealed distinct genomic loci where transcription activity patterns of TSS within are highly correlated. (**Figure 2.8**) The extent to which transcription activity is correlated among many adjacent individual TSS was unexpected. Our results revealed that not all adjacent and coregulated regulatory elements contain TLR4 signaling effector transcription factor motifs such as NF-κB or AP-1. This suggests that these sites might be activated bymechanisms other than direct transcription factor action,such as communication between the regulatory elements found within close proximity to each other. (**Figure 2.7**)

2.5 Discussion

Here we report the versatility of csRNA-seq to identify regulatory cascades within our model system. This cascade begins with the activation of NF-κB motif containing regulatory elements, followed by a secondary wave of AP-1 motif-containing CREs, followed by a tertiary wave of IRF-enriched CREs. The use of csRNA-seq to profile nascent transcription dynamics at both enhancers and promotors allows base-resolution identification of regulatory elements responding directly to the stimulus. We also observe a loss of transcription activity at groups of TSS not enriched for NF-κB, AP-1 or IRF but are enriched for ETS sites, suggesting they are not targeted by our immune signaling pathway. Together this transcriptional response within our model system is robust with the first cascade of TSS being induced within 10 mins of LPS exposure. This first cascade of responding TSS were enriched for NF-κB motifs, which verifies initial timing estimates of NF-κB TFs entering the nucleus within 10 mins after stimulation (Ferreiro & Komives, 2010). These results also show how while the first transcriptional burst is predominantly NF-κB -dependent induction, another, stronger transcriptional burst containing primarily AP-1 sites along with NF-κB sites provides a glimpse into co-regulatory dynamics between these TFs. This robust induction verified a successful activation of the TLR4 pathway with our model system and data. This verification permits further study into transcription induction, to decipher what genomic attributes are associated with transcription induction precursors. At the forefront we observe the induction of NF-κB containing sites as a likely genomic precursor leading regulatory induction dynamics within LPS-treated macrophages. Along with NF-κB sites we also note the induction of TSS that lack NF-κB or AP-1 motifs that are located within close proximity of NF-κB-containing sites (~15,000 bp). To further analyze this phenomenon, we focused on a prototypic LPS response gene, TNFAIP3, which is associated with

a proportion of induced TSS that did not contain NF-κB or AP-1 transcription factor binding motifs. This leaves one to question as to how these TSS are induced and whether and by what means they are participating in the activation of the TNFAIP3 gene. We hypothesized that these regulatory elements, located within proximity to induced NF-κB-containing TSS are not directly recruiting TFs but rather are participating in and are activated due to their proximity to other TSS.

This would suggest a model where distal REs may act as the direct means of recruiting, modulating and most importantly maintaining an optimal concentration of transcription machinery within their genomic loci and by so doing indiscriminately influence the transcription of other adjacent REs and gene bodies within the created TF rich environment. (**Figure 2.9**) The fact that we see all REs within a domain induced, including those that do not have stimulation-associated transcription factor binding motifs (TFBMs), suggests communication with other REs is relatively non-specific and based only on their linear proximity along the genome (and potentially their spatial proximity). The fact that shared induction applies equally to promoters and enhancers implies these elements might be equal players in this process, underscored by the fact that their motifs and chromatin features are largely indistinguishable (Core et al., 2014). Collectively, the regulatory elements within a domain may recruit and maintain a high local concentration of transcriptional regulators, an idea that might work in conjunction with the downstream impacts of KLA signaling and the formation of phase separated condensates comprised of transcriptional proteins. These genomic environments of localized shared induction dynamics, demonstrated by the activity at the TNFAIP3 gene locus within KLA-treated macrophages. We identify these genomic regions of intense transcriptional activity, exhibited by the TNFAIP3 locus, as cis-regulatory domains (CRDs). We hypothesize that the CRDs are the functional result of one or few CREs driving the recruitment of transcription factors whose recruitment influences transcriptional

activity at CREs within the domain. (**Figure 2.9**) To address the veracity of our proposed model the following chapters will carefully deconvolute the transcription signals as well as genomic attributes identified within these sites of correlated activity. By leveraging csRNA-seq and various genomic perturbation methods we have the means to directly test the potential mechanisms by which transcription is initiated and modulated at the genomic level.

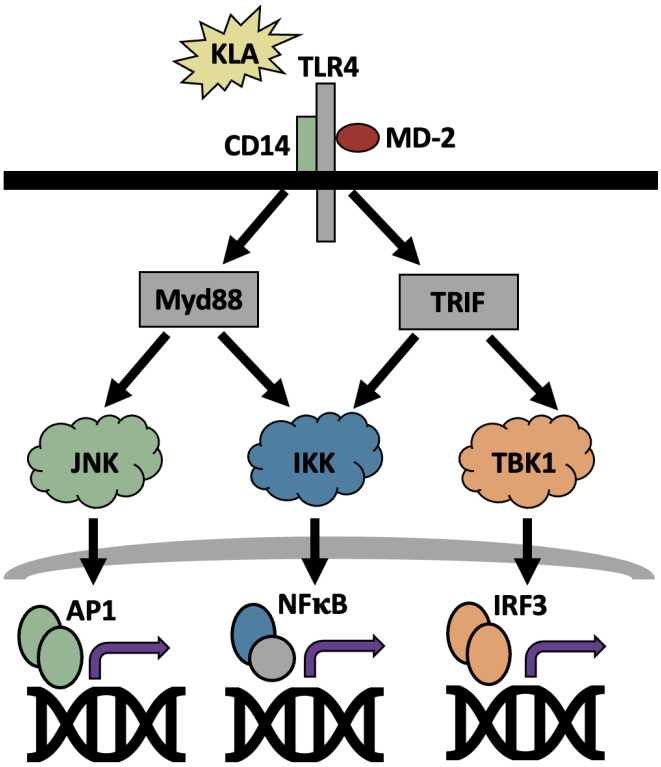

Figure 2.1 Overview of the TLR4 pathway. KLA activates the TLR4 signalling cascade leading to direct transcription factor – DNA interactions.

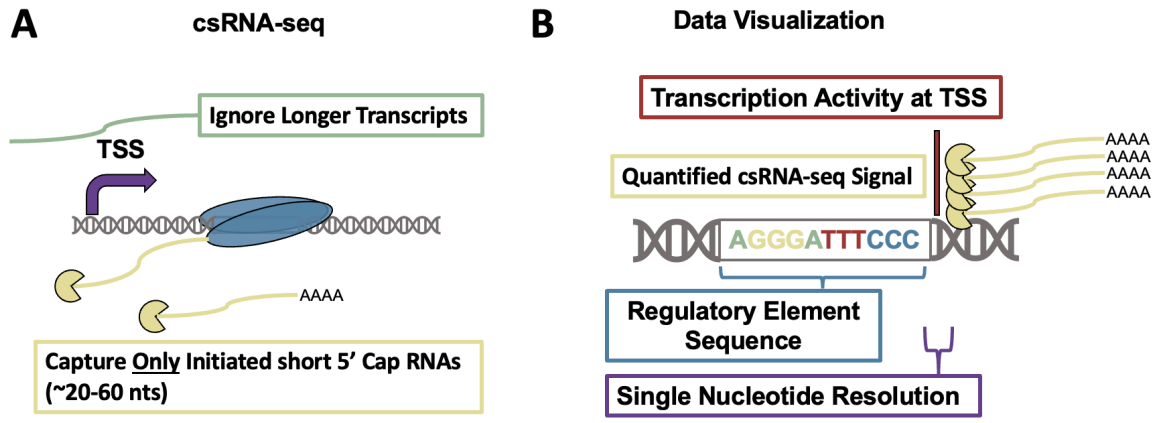

Figure 2.2 Diagram of csRNA-seq. A) csRNA-seq is the capture of short, initiated RNA transcripts formed as RNA Polymerase II begins transcription. Longer stable RNA transcription are ignored. B) The quantification of csRNA-seq reads (yellow) reveals the transcription activity (red) at a single nucleotide resolution (purple) which allows the identification of sequence elements (blue) associated with the transcription activity observed.

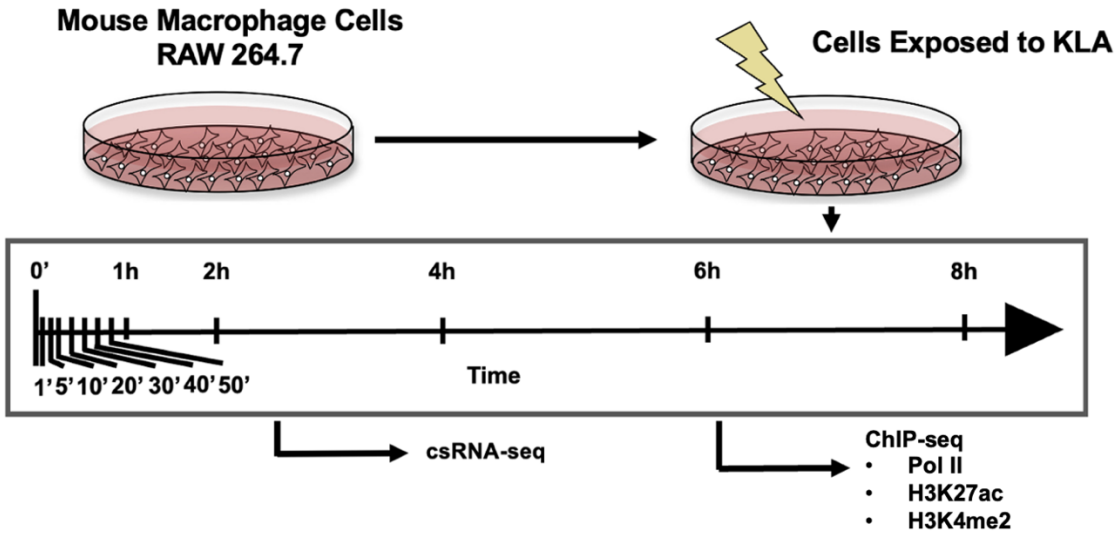

Figure 2.3 Overview of experimental design, the model dataset is be comprised of csRNA-seq samples taken from macrophages cells that have been exposed to KLA for a given time interval.

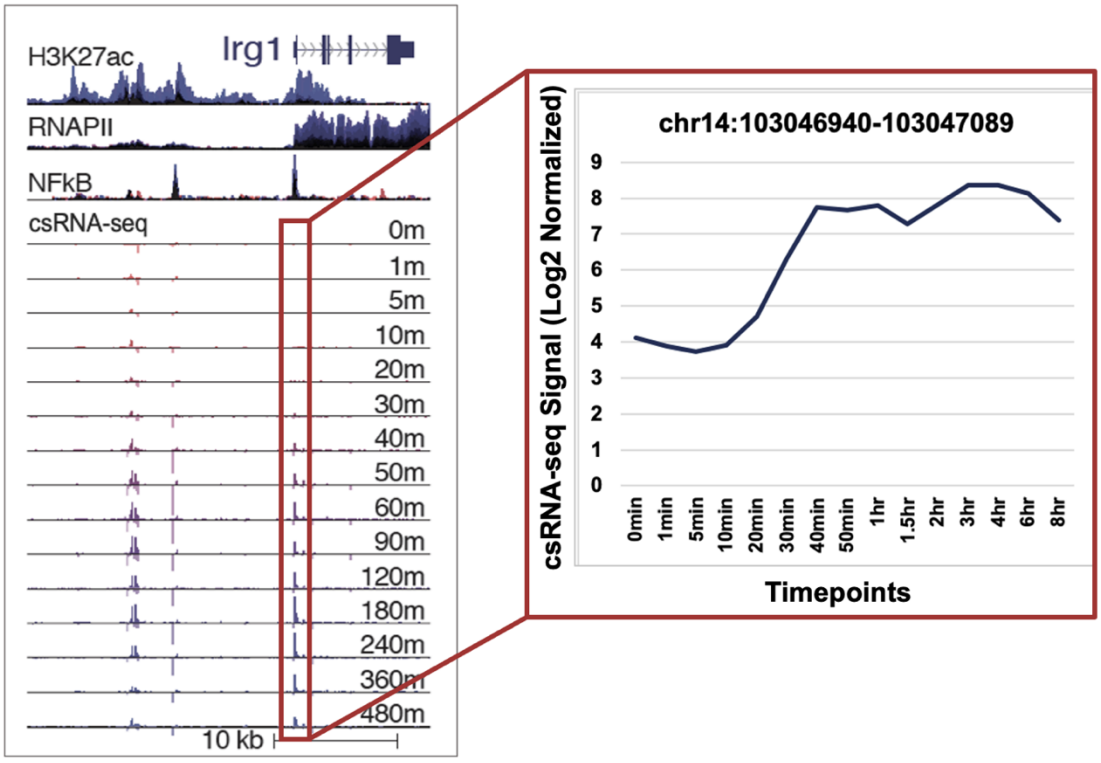

Figure 2.4 Visualization of csRNA-seq data as collected at individual timepoints. These timepoints can be plotted to demonstrate its transcription activity profile.

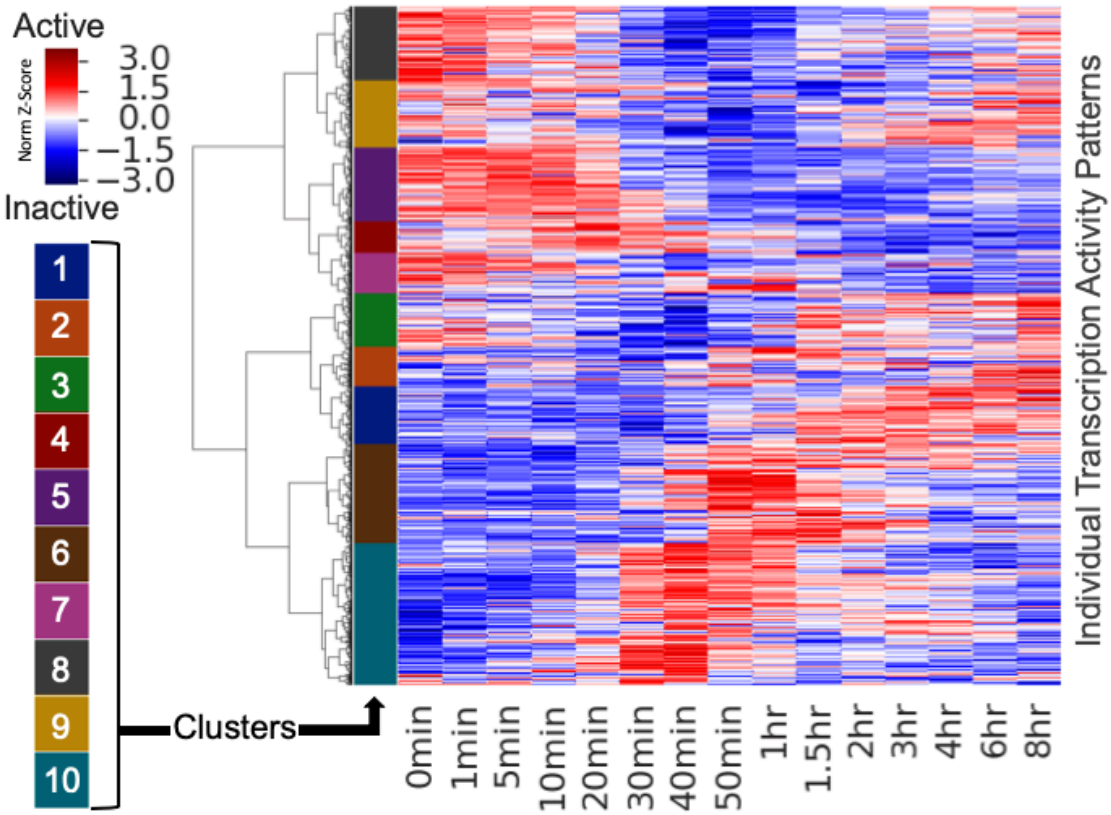

Figure 2.5 Cluster map of hierarchically clustered activity patterns of transcription initiation at all transcribed regulatory elements genome wide. The data within the heatmap represent z-scored, rlog normalized csRNA-seq signal at csRNA-seq positive genomic locations (rows) at indicated timepoints (columns) after KLA treatment. [N = 69927 sites of transcription initiation]

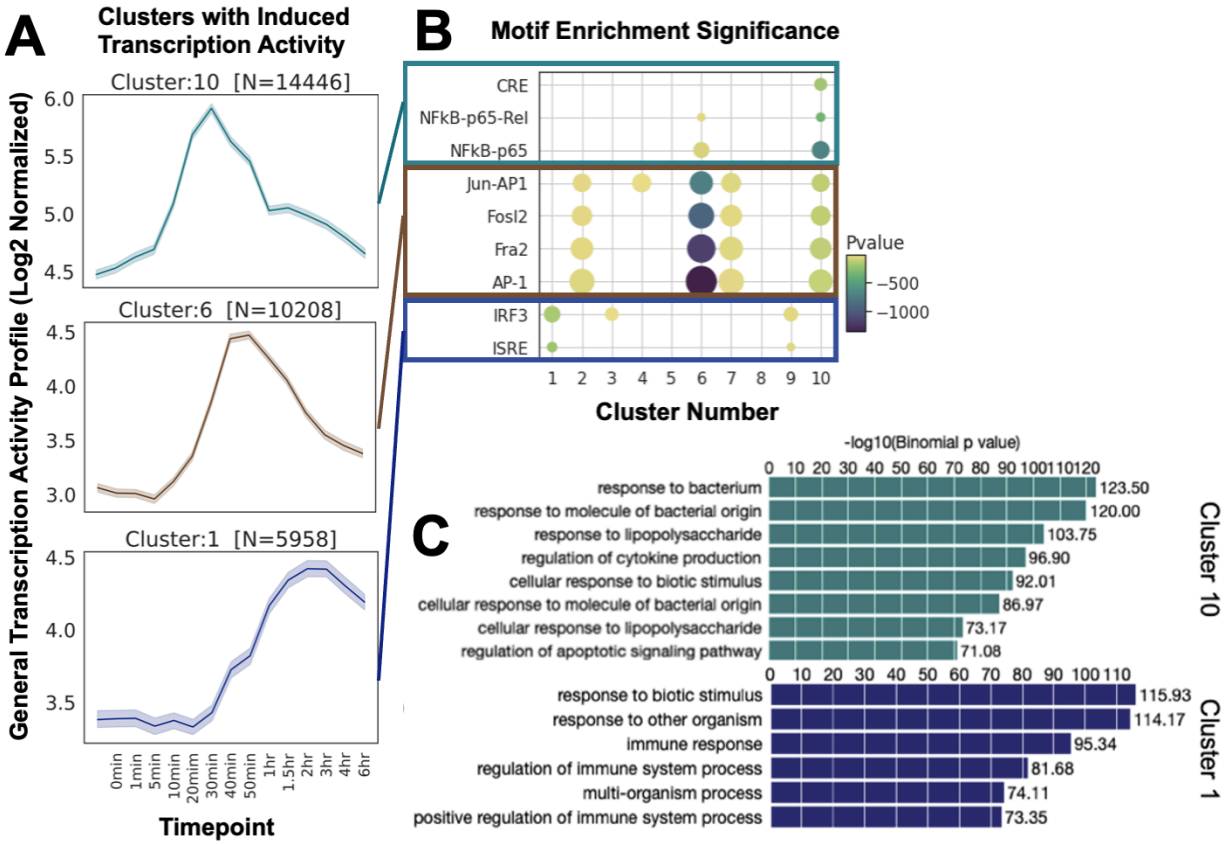

Figure 2.6 Overview of results. A) Distinct transcription activity patterns for clusters 1, 6 and 10. B) Motif enrichment of selected motifs across all transcription activity pattern clusters. C) GO analysis (Biological Processes) results for clusters 1 & 10.

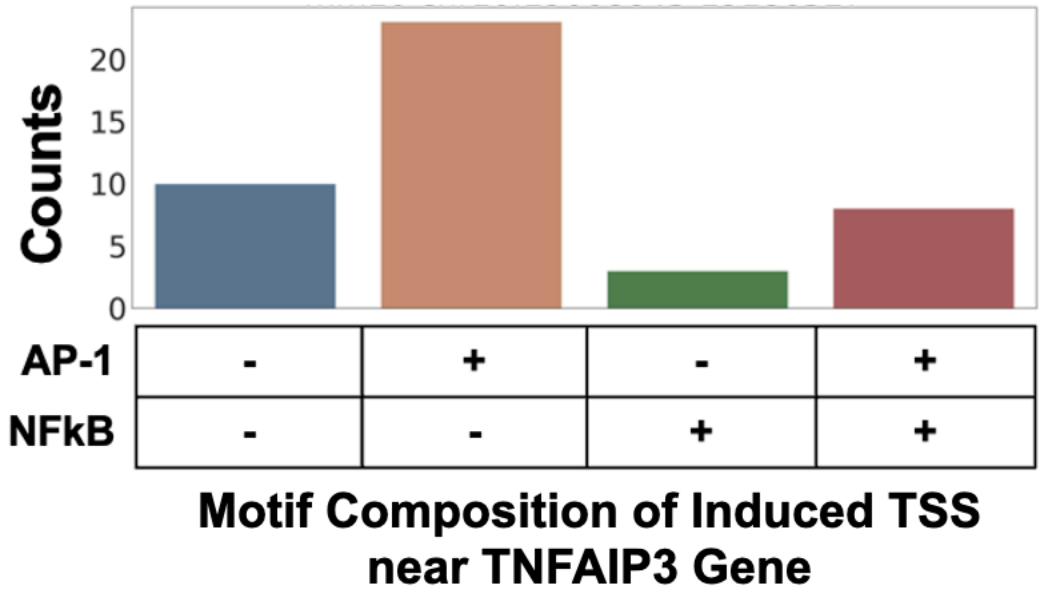

Figure 2.7 Motif composition of correlated TSS found at the TNFAIP3 locus: chr10:19008643-19186527.

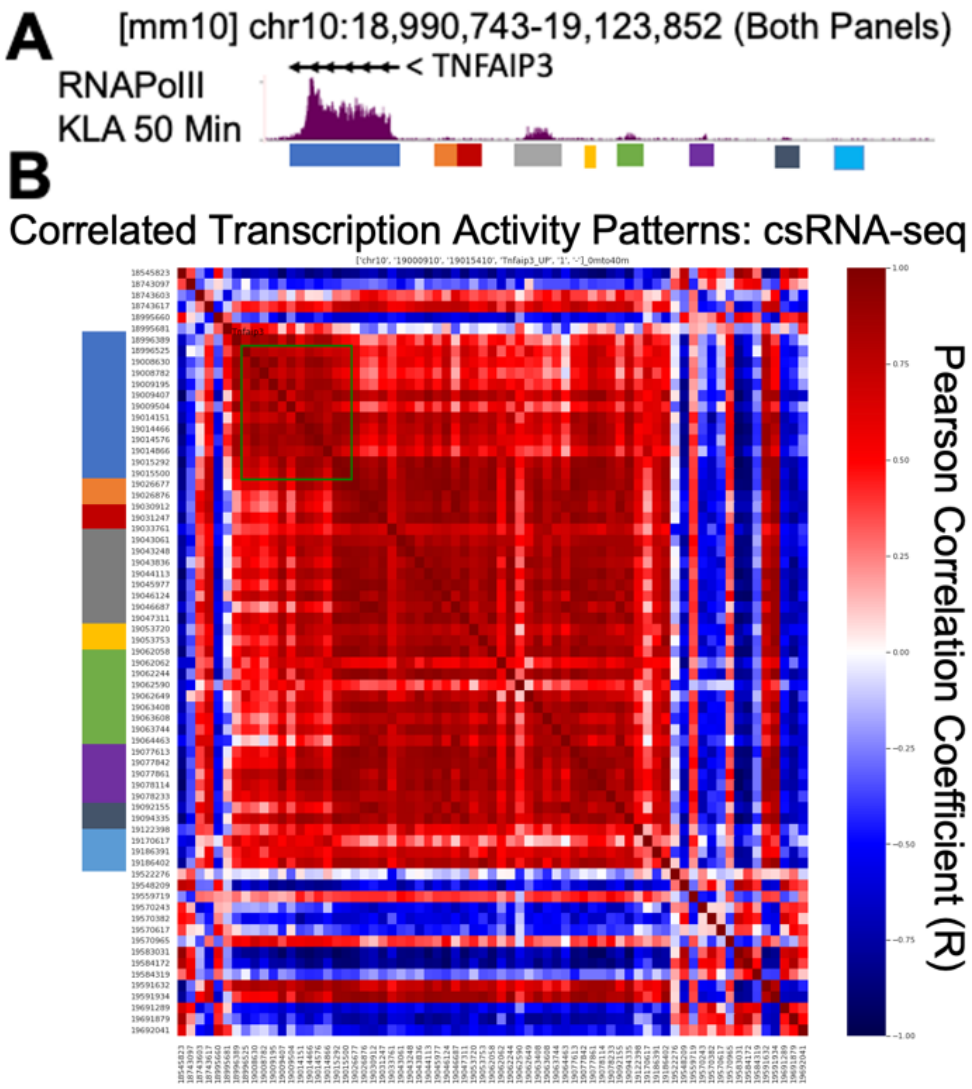

Figure 2.8 Transcription activity near TNFAIP3. A) ChIP-seq (Pol II) 50 minutes after KLA stimulation showing elongation at the gene body and distal regulatory elements. B) Heatmap of correlated transcription activity patterns as detected within the csRNA-seq time course.

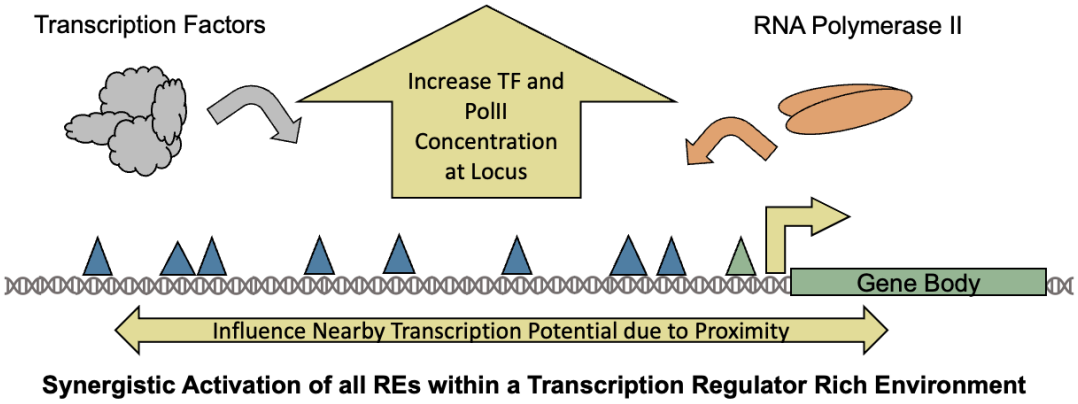

Figure 2.9 Overview of the proposed model tested within this thesis.

Acknowledgements

Chapters 1,2,4,5,6 represent research intended to be published together pending further insight provided by a Dual-MPRA approach described in Chapter 6. Hillary, R., Guzman, C., Heinz, S., Benner, C. The dissertation author will be the primary investigator and author of this paper.

Chapter 3 Induction Timing Characterization of Localized Enhancers and Promoters

3.1 Abstract

The development of csRNA-seq permits the capture of stable and unstable transcripts generated from the initiation of transcription at regulatory elements genome-wide. The methodology employed by csRNA-seq also permits capture of these transcripts within very short time intervals, forgoing the lengthy time intensive steps such as crosslinking requirements within ChIP-seq. Identifying the precise timing of transcription of initiation at individual regulatory elements has been a crucial question in deconvoluting regulatory transcription activity signals, especially as it relates to regulatory elements located proximal (promoters) or distal (enhancers)

relative to gene bodies. It is believed that transcription at enhancers occurs before promoters, which is a key observation supporting promoter-enhancer looping (Arner et al., 2015). Using improved methods, such as csRNA-seq along with a greatly improved induction calculation that limits biases from sequence stability and signal intensity we observe that the induction of transcription initiation at activated enhancers and their regulated promoters is simultaneous. With enhancer/promoter induction being nearly simultaneous it raises the question of what genomic properties of a regulatory elements, beyond their proximity to a gene body, are more accurate attributes to assess if a regulatory element would precede another in transcription initiation. Considering this we focus on the composition of the regulatory element rather than its proximity to gene bodies for genomic attributes driving their initiation.

3.2 Introduction

The earliest theory that attempts to characterize enhancer/promoter mediated gene expression and modulation is the looping theory (Arner et al., 2015). This theory describes gene expression as being the direct result of enhancers, after recruiting transcription factors (TFs), looping to physically bind a promoter to mediate the recruitment of additional transcription machinery, including RNA polymerase II to transcribe the gene body itself (Bulger & Groudine, 1999). For enhancers to exhibit initial transcription induction it would infer that they are actively the early target for transcription factors and regulatory mechanisms orchestrating recruitment, initiation and modulation of transcription. For promoters to be induced after enhancers could indicate that their regulation and modulation is dependent on enhancer activity. If induction timing between regulatory elements differs from the current model, in that enhancer induction is not always preceding the induction of promoters, then the looping theory may not address true dynamics governing enhancer/promoter interactions.

Previous studies have attempted to resolve the relative induction timing between enhancers and promoters, genome wide. Arner et al. demonstrated using CAGE that enhancers are induced in advance of promoters after both changes in cell fate and cell state. However, the recent development of new profiling techniques allows us to revisit this question with unprecedented accuracy and resolution. csRNA-seq allows for high resolution time courses that capture nascent-like transcriptional activation at both enhancers and promoters. Other methods of measuring transcription activity either target specifically stable transcripts (RNA-seq) or target all transcripts with a heavy signal bias toward stable transcripts (CAGE-seq). This signal bias favoring the capture of stable transcripts is problematic in that stable transcripts (e.g. messenger/promoter RNAs) will appear to increase in level over time (accumulate) relative to unstable transcripts (e.g. enhancer RNAs), even if they are transcribed at similar levels. The use of csRNA-seq (or other nascent methods) is ideal for measuring transcription dynamics within tight time intervals, reducing the need to computationally account for transcript stability required for longer interval time series data.

By leveraging csRNA-seq to revisit the question of enhancer and promoter induction timing we found enhancers generally do not precede promoters within our model system. By examining each induced and localized enhancer/promoter pair genome wide we found the mean induction time between these pairs to be near simultaneous.

3.3 Methods

3.3.1 csRNA-seq: Preferred Method of Transcription Induction Profiling

With the multitude of RNA-profiling techniques available it is important to consider their caveats and limitations as these biases can interfere with careful interpretation and deconvolution of transcriptomic signals. In order to effectively assess a question such as the relative induction

timing of enhancer and promoter there are a few elements to consider such as RNA stability, TSS positional resolution and methodological limitations that limit short interval profiling approaches. RNA stability is a crucial caveat to be considered when directly comparing and profiling regulatory elements both distal and proximal to gene bodies, as stable transcripts are more prevalent at promoters as they result in stable gene transcripts while enhancer RNA are typically unstable and targeted for degradation. While most RNA profiling techniques are successful in capturing stable RNA products, csRNA-seq specifically targets newly initiated RNA polymerase II transcripts which accurately represent both the eventual stable and unstable RNA species (Duttke, Chang, Heinz, & Benner, 2019). The relatively equal capture of both stable and unstable transcription products permits the accurate capture and comparison of proximal and distal transcription activity dynamics, forgoing the inherent bias toward stable transcript capture that most RNA profiling techniques exhibit that make direct enhancer/promoter transcription comparisons problematic.

In terms of positionality, csRNA-seq permits the identification of the specific bp from where such transcription initiation activity is originating, such that individual TSS comparisons, rather than broad genomic loci, are possible. Finally, because csRNA-seq is feasible in any system where only total RNA can be extracted and forgoes time intensive protocol requirements it is conducive to be used in very short interval time series experiments.

3.3.2 Inclusion of Tighter Interval Time Series

While the initial research presented here focused on TLR4 pathway induction dynamics and leveraged csRNA-seq data from multiple timepoints within KLA-treated macrophages, a shorter-interval time series dataset was analyzed to verify enhancer/promoter timing findings. It is known that key transcription factors, namely NF-κB, enter the nucleus within 10 mins after KLA detection (Ferreiro & Komives, 2010). This rapid translocation of a transcription factor to the

nucleus in such a short timeline enables comparison of transcription dynamics at both enhancers and promoters. A time series dataset was produced after collecting samples every 2 minutes within the initial 20 mins of KLA treatment in macrophages. (**Figure 3.1**) This dataset was used to complement our original 10-min interval time series used to examine primary, secondary and tertiary TLR4 pathway transcription activation dynamics in Chapter 2.

3.3.3 Transcription Induction Metric

Given the dynamic nature of time series data the challenge is to select a specific metric that represents a timepoint in which induction is occurring at a given Transcription Start Region (TSR). We determined the time at which a 50% increase was detected between baseline signal (0 min) and the relative max signal was the least biased from differences in signal intensity. (**Figure 3.2**) We will refer to this metric as the 50^{th} percentile metric. As a direct comparison and to demonstrate how other measures may have inherent biases due to signal intensity as well as RNA stability, we also applied the use of center of mass index (CMI) within our results (**Figure 3.3**) (Arner et al. 2015).

3.3.4 Enhancer Promoter Pair Selection Criteria

Because of the dynamic nature of transcription within our model system we focused our direct enhancer and promoter induction timing comparisons on enhancers within +/- 100,000 bp of promoter sites. We applied this spatial constraint and utilized a metric we will refer to as Shift, this implementation of this metric is adopted from previous enhancer/promoter timing studies (Arner et al., 2015). The Shift metric represents the induction time measurement of the promoter minus the induction metric of an enhancer within +/- 100,000 bp of the promoter. A negative shift metric represents enhancer induction preceding promoter induction, meaning the paired enhancer is transcriptionally active before its promoter counterpart. (**Figure 3.4**)

3.4 Results

3.4.1 Signal Bias Among Select Induction Metrics

Arner et al. reported that enhancers are induced before promoters, and among the studies that report this, CAGE-seq is typically used to measure induction timing by calculating the center of mass (**Figure 3.3**) (Arner et al., 2015). This center of mass calculation leverages multiple time intervals to generate an index which is referred to as CMI which became the basis of comparison between our revisiting of enhancer/promoter induction timing. CAGE has a detection bias toward stable RNAs compared to csRNA-seq which similarly detects nascent short RNAs, for unbiased unstable/stable RNA detection and profiling. Firstly, the CMI metric was applied to two csRNAseq time-series datasets: a longer interval csRNA-seq timeseries dataset with 10-min intervals as well as a tight interval dataset with 2-min intervals. Applying CMI to RNA-seq and CAGE-seq data as described by Arner et al. we found that enhancers appear to have an earlier induction relative to promoters. However, using CMI within our long interval csRNAseq dataset we found little difference between enhancer/promoter timing. (**Figure 3.5**)

Upon further examination of the CAGE-seq TSS data from Arner et al., we found that CMI exhibits signal bias toward capturing stable RNA species, which makes determining accurate enhancer transcription activity difficult. This signal bias can be attributed to RNA stability. When CMI is applied to the entire time series dataset as demonstrated by Arner et al., the signal decay at promoters is prolonged and is causing the "delay" in promoter induction relative to enhancers. The use of CMI to measure "induction" across an entire time series data profile along with bias CAGE-seq toward stable mRNAs have led to the erroneous conclusion that enhancers are induced before promoters. (**Figure 3.6**) If CMI is limited to an induction window, defined as the interval from timepoint 0 min to the timepoint of maximal induction, we find that the timing between enhancers

and promoters within csRNA-seq and CAGE-seq shows little difference. Using CMI to first examine timing we found that within CAGE-seq the enhancers had a mean CMI of 88 min while the promoters returned a CMI of 95 minutes. While RNA-seq and CAGE-seq differ in approaches and target RNA species they share a similar bias toward stable RNA transcripts.

To bridge the gap between the CAGE-seq data (Human + BMDM) and our model system (Mouse RAW264.7 cells + LPS) we analyzed RNA-seq time series data generated in RAW cells treated with LPS (10 min intervals). Within this RNA-seq timeseries we found that enhancers had a mean CMI of 77 minutes while the promotors had a mean CMI of 86 minutes, replicating the enhancer preceding promotor dynamic reported by Arner et al., However within the longer time interval (10-minute) csRNA-seq dataset, we found a mean CMI of 99 min was shared between enhancers and promoters. Within the CAGE and RNA-seq time series we found the Shift to be -10mins and -20mins, respectively, while the csRNA-seq dataset returned a Shift of -2 min. While the Shift metric is used to directly compare enhancer/promoter induction dynamics relative to one another, CMI can be problematic when applied to the entire time series dataset. We found that the CMI of a time series rarely marked timepoints during the window of increased transcription activity within a given kinetic profile but rather returned timepoints during the decay part of the transcription signal curve.

3.4.2 Enhancer Promoter Induction Timing Analysis

Using the 50^{th} percentile metric along with the Shift metric to compare enhancer and promoter pairs we found that within our longer interval (10-min interval) csRNA-seq time series the 50^{th} percentile Shift was +2 mins, showing a slightly earlier promoter induction relative to enhancers. (**Figure 3.6 & Figure 3.7**) Given this 2 min shift is shorter than our 10-min sampling intervals and given the rapid and robust induction of the TLR4 pathway upon endotoxin

recognition we applied the same 50th percentile metric Shift to the shorter-interval (2-min interval) csRNA-seq time series. Within the short interval time series, we found an average Shift of -1 min between enhancers and promoters which again is between the 2 min interval sampling approach employed. These results demonstrate that enhancer and promoter induction timing is not as clearly skewed between one type of regulatory element vs the other and suggests their timing is near simultaneous down to a 2min time series resolution for the given model system. (**Figure 3.8**)

3.4.3 Characterization of Transcript Stability Influence on Induction Timing Interpretation

To reconcile our csRNAseq + 50th percentile results with previous CAGE-seq + CMI findings it is important to consider RNA stability and its role to influence metrics measuring induction timing. When analyzing RAW cells exposed to LPS and examining the robust transcriptional dynamics within a time series experiment it becomes apparent that each TSS will contain a timepoint where transcription is maximized. Leveraging this timepoint of maximized transcription activity we examined both the induction moment and the decay moment, the former being the ramping up of transcription activity at a TSS up to the max and the decay representing the tapering off of transcription activity. (**Figure 3.9**) These two-time intervals, induction and decay, represent a crucial insight into RNA stability and its effect on interpreting enhancer/promoter transcription dynamics. Using the point of maximum transcription activity as a relative timepoint we can leverage the 50th percentile metric to examine both the induction and decay at each TSS. The metric used will be referred to as Induction and Decay. The induction is calculated as the time of maximal transcription activity minus 50th percentile of the induction. The Decay is calculated as the 50th percentile of the TSS decay interval minus the time of maximal transcription activity.

Using this approach, we report promoters exhibit a lengthy decay in transcription activity within our CAGE and RNA-seq datasets. In contrast transcription activity measured at enhancers within the CAGE and RNA-seq datasets exhibited a short decay. When enhancer and promoter transcription activity signal decay was examined within csRNA-seq the mean decay times were comparable. The lengthy transcription decay at promoters relative to enhancers exhibited within the CAGE and RNA-seq datasets indicates that differential RNA stability of promoter (i.e., gene) and enhancer transcripts is a confounder when interpreting and comparing their transcription dynamics.

3.4.4 Effects of Specific Genomic Characteristics on Enhancer/Promoter Induction Timing

With our current results suggesting induction timing between enhancers and promoters to be near identical within our model system we set out to identify if specific genomic attributes could perhaps cause a divergence between induction timing. (**Figure 3.10**) While still dividing regulatory elements into enhancer and promoter groups we added additional dividing criteria and calculated 50^{th} percentile shift metrics to identify which characteristics affect induction timing. Focusing specifically on the short-interval csRNA-seq time series dataset we first examined if regulatory elements containing canonical motifs, in this case specifically NF-κB transcription factor binding motifs. NF-κB sites being one of the primary targets of the TLR4 pathway and are induced before sites lacking that specific motif. Between the NF-κB motif-containing and lacking loci we observed little difference in induction timing. Focusing on specific regulatory elements were NF-κB is detected to be bound using ChIP-seq, we again found no significant differences between NF-κB bound and NF-κB non-bound sites. While our model system focuses on LPS-stimulated mouse macrophages, a system that focuses on highly conserved innate immune responses rather than prolonged systems such as cell fate we decided to focus on another key

relatively dynamic genomic attribute: the remodeling of chromatin. (Hargreaves et al., 2009) Using ATAC-seq and annotating individual TSS with detectable changes to DNA accessibility (increasing accessibility, decreasing accessibility and no change to accessibility) we assessed if changes in relative accessibility could influence induction timing within our group of regulatory elements. While more dynamic as to potential changes in induction timing relative to other attributes discussed thus there was little difference between each group (Figure 3.10). For the sites that reported increases of accessibility relative to unchanged/unknown accessibility status showed slight shifts toward "earlier induction" for the enhancer/promoter group experiencing the increased accessibility. While this observation is not statistically significant due to the low sample sizes for these individual groups, this trend is clearly visible. Finally, we examined if direct loci-loci interactions within the +/-100,000-bp window could have an influence in their relative induction timing and annotated loci as to if physical interaction was detected using PLAC-seq. Again, we detected little shift in induction timing between these physically looped or non-looped enhancer/promoter groups. While slight shifts are detectable the relative significance between groups suggests the conclusion is this: using a short-interval (2 min) time series examining LPS-treated mouse macrophages the timing between localized (+/- 100,000 bp) enhancer and promoter pairs shows little difference in induction timing.

3.5 Discussion

Previous reports suggest that transcription at enhancer sites precedes transcription at promoter sites, which would align with the enhancer-promoter looping theory of enhancer function. Our results demonstrate a very different dynamic where, when examined genome wide and specifically examining stimulus-based induction patterns, the induction timing is near simultaneous. This simultaneous induction can also be interpreted such that either enhancers or

promoters could be induced before the other within individual loci and their induction potential isn't defined by this distinction. This demonstrates that induction between either enhancers or promoters is indiscriminate and other factors are the intended targets within these loci and domains rather than explicitly their relative proximity to gene bodies. Similar correlated enhancer/promoter dynamics and phenotypes that contradict the conclusions by Arner et al. that enhancers are transcribed before their target promoters have been observed in other stimulus responses. This includes T cell activation (Michel et al., 2017) and activation of estrogen receptor 1 in MCF7 cells (Hah et al., 2013). It will be interesting to extend these studies to changes in cell fate such as cell differentiation, which represents key future work to establish using our induction metric and transcription measurement method (csRNA-seq).

Within these results we also demonstrate potential risks involved with interpretation of transcription dynamics based on RNA stability and specific RNA species targeted through assays. We directly compared our approach using csRNA-seq with another approach that leveraged CAGE employed by Arner et al., 2015 examining multiple time series datasets. Within our analyses we used our mentioned 50^{th} percentile metric while the applied method within Arner et al., 2015 was the center of mass index (CMI) metric. We noted that when CMI was applied to both csRNA-seq as well as CAGE the CMI was highly biased toward sites with lengthy induction decay. This prolonged decay in induction signal was specific to CAGE as well as RNA-seq which represents the disparity of RNA stability between enhancers and promoters within these capture methods. This RNA stability bias is further highlighted in that the use of CMI was applied to the entire transcription activity profile, including all timepoints beyond 300 mins after LPS exposure. This resulted in CMI values being reported specifically within the LPS + BMDM dataset to be reported long after the transcription's induction moment and maximal induction timepoint. These CMI

values reported for LPS + BMDM simply do not represent the timepoint where induction is occurring and coupled with RNA stability prolonging detection relative to unstable RNA species, will hinder result interpretations. It was encouraging to see CAGE and csRNA-seq both report 50th percentile metrics showing similar induction timepoints but also similar enhancer promoter induction dynamics. The use of a very short interval (2m) short timeseries represents a "good faith" examination of our results to verify down to a short resolution that induction dynamics are similar.

While we apply the distinction of enhancer and promoter to individual regulatory elements due to characteristics such as proximity to gene bodies as well as genomic attributes such as core promoter elements; their similar induction timing suggest the initiation within individual loci isn't specific targeting one or the other. This is an important distinction in that induction potential at individual loci seems more weighted toward regulatory element composition rather than simply proximity to gene bodies. By limiting our analysis to early induction patterns, we confirmed an enrichment of NF-κB transcription factor binding motifs and other potent activators such as AP-1 as the principal sites of induction of the TLR4 pathway. Our work represents continued focus on these genomic attributes being the key influencers of transcription factor recruitment and induction of loci-wide transcription. However, a limitation to this interpretation lies with the observation that all relatively simultaneously induced loci within our analysis contain these canonical TLR4 associated transcription factor binding motifs. While we assert transcription induction isn't limited to singular enhancer/promoter interactions, we add evidence that enhancer/enhancer enhancer/promoter interactions could be at play and their induction potential and activity could be communicated to nearby adjacent regulatory elements. This mechanism being in line with the current regulatory model of liquid-liquid phase separation (LLPS) (Sabari et al., 2018).

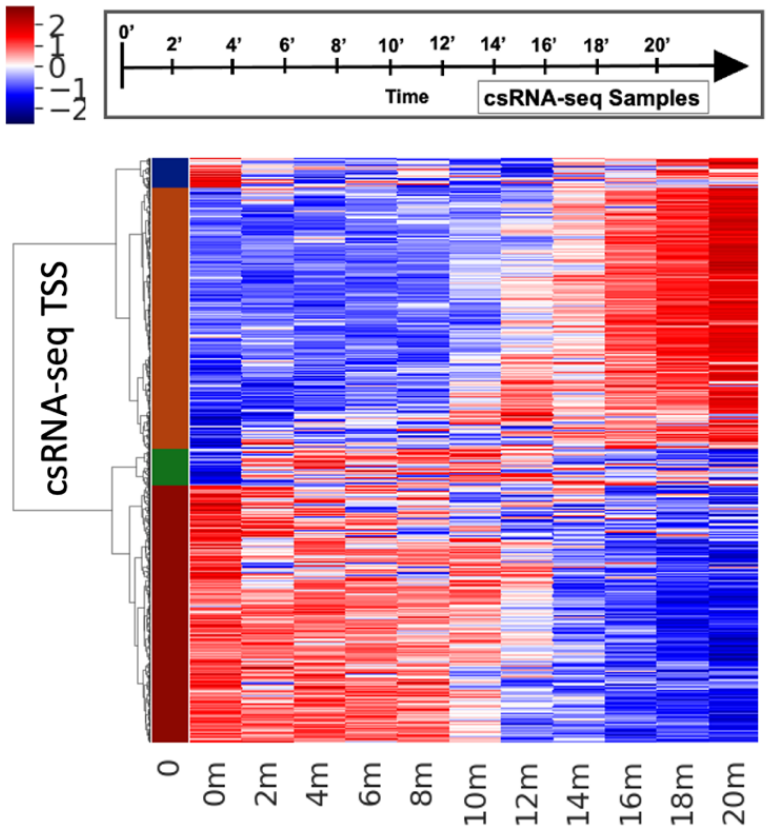

Figure 3.1 Cluster map of hierarchically clustered activity patterns of detected transcription initiation at all regulatory elements genome wide within a short interval time series.

A	$time_{beforeTargetSignal}$	$f(t) = time_{targetSignal}$
B	$time_{targetSignal}$	$t = signal_{targetSignal} = \frac{signal_{maxTime} + signal_{minTime}}{2}$
C	$time_{afterTargetSignal}$	$f(t) = \frac{time_{afterTargetSignal} - time_{beforeTargetSignal}}{signal_{afterTargetSignal} - signal_{beforeTargetSignal}}(t - signal_{beforeTargetSignal})$
D	$signal_{maxTime}$	
E	$signal_{afterTargetSignal}$	
F	$singal_{targetSignal}$	
G	$signal_{beforeTargetSignal}$	
H	$signal_{minTime}$	

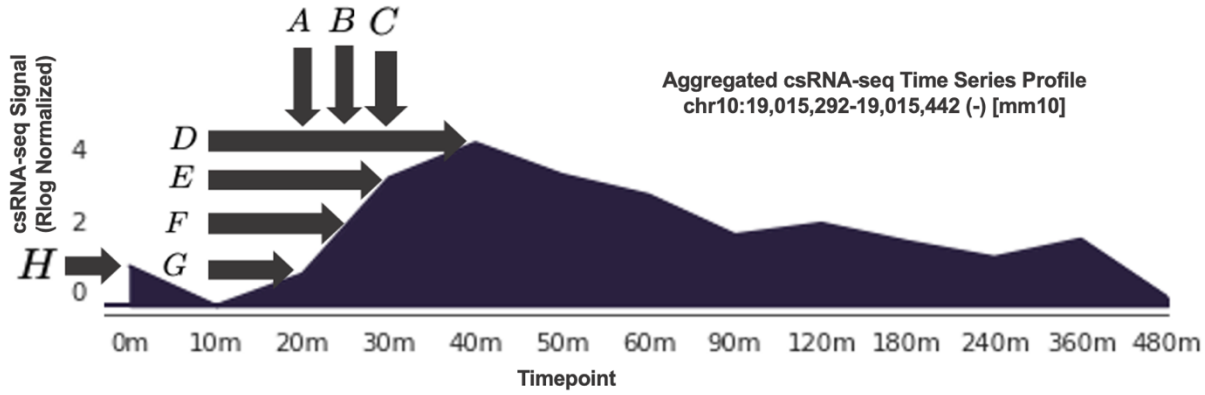

Figure 3.2 Description and diagram of the 50th percentile metric calculation using both signal values and time points.

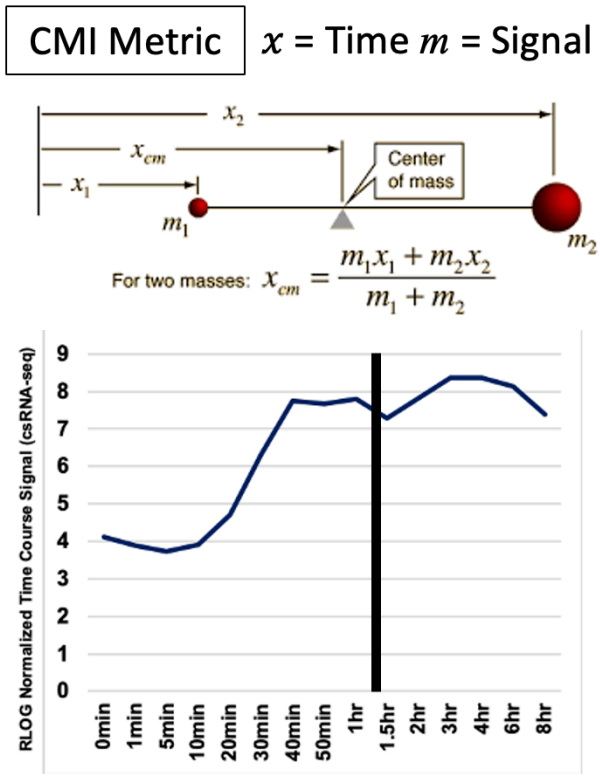

Figure 3.3 Center of mass calculation as applied to time series transcription data

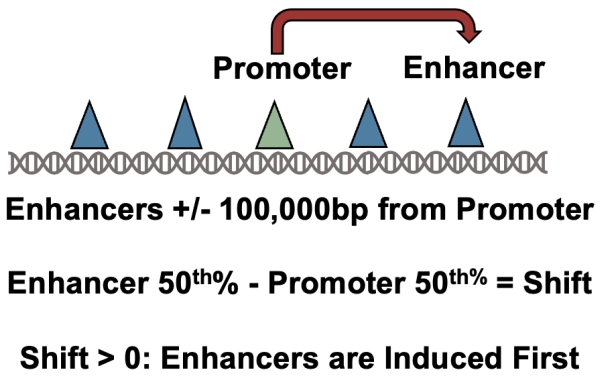

Figure 3.4 Schematic of how the Shift metric is calculated.

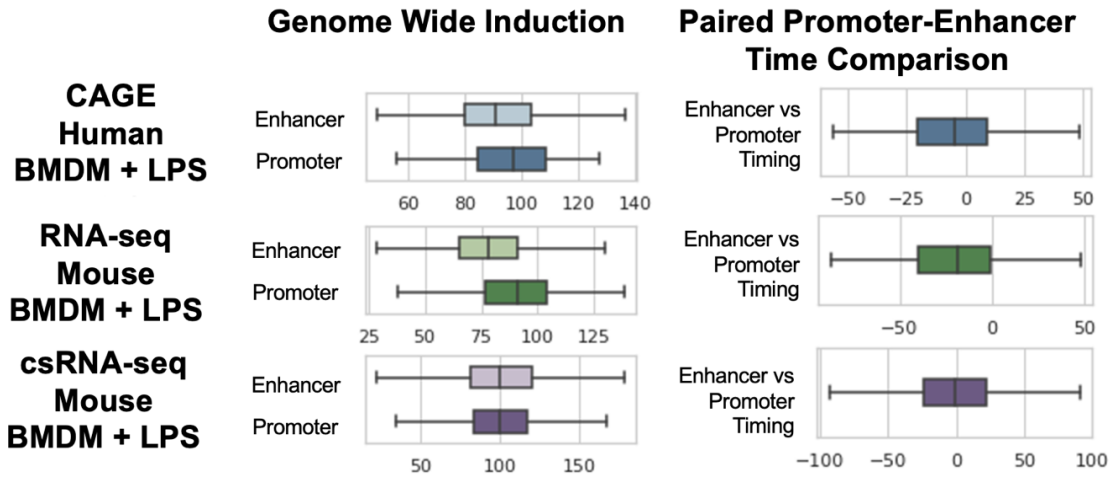

Figure 3.5 Center of Mass Index (CMI) and Shift results for CAGE, RNA-seq, and csRNA-seq datasets.

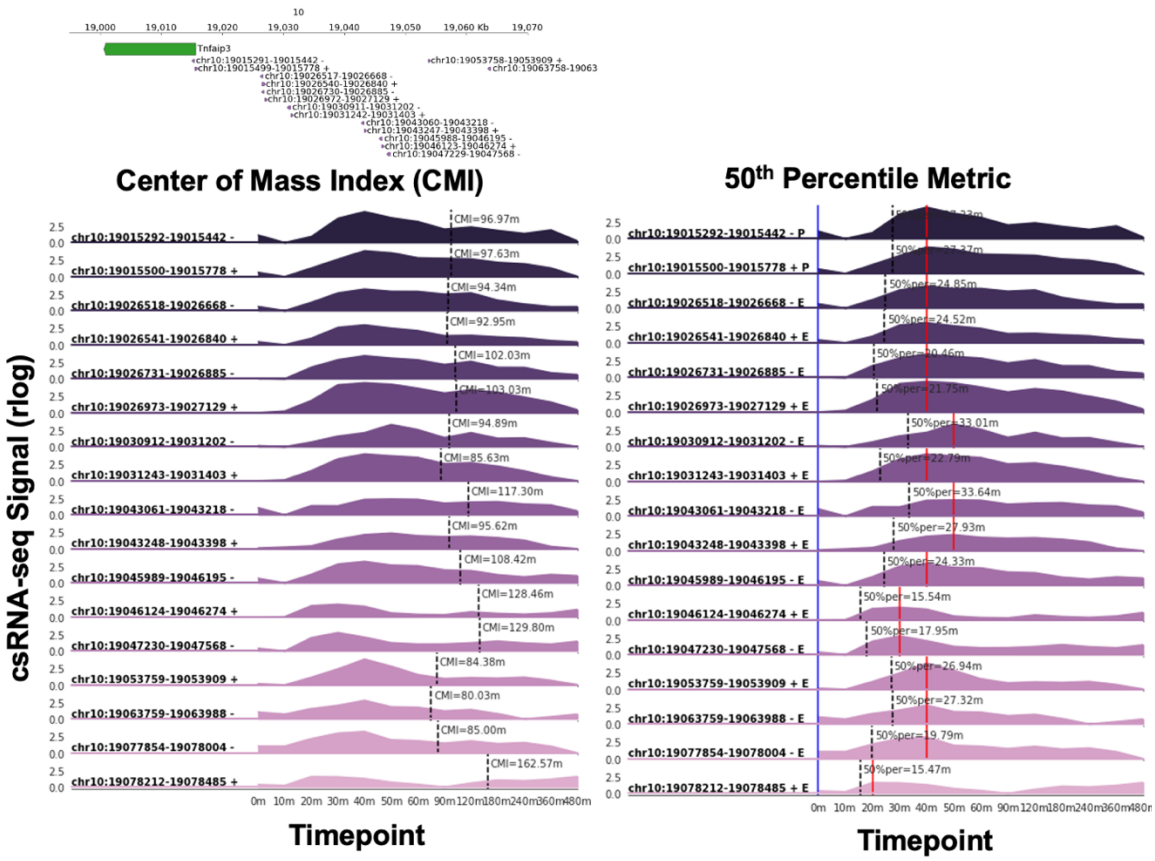

Figure 3.6 Direct comparison of CMI values vs 50th percentile metric values within the same genomic locus.

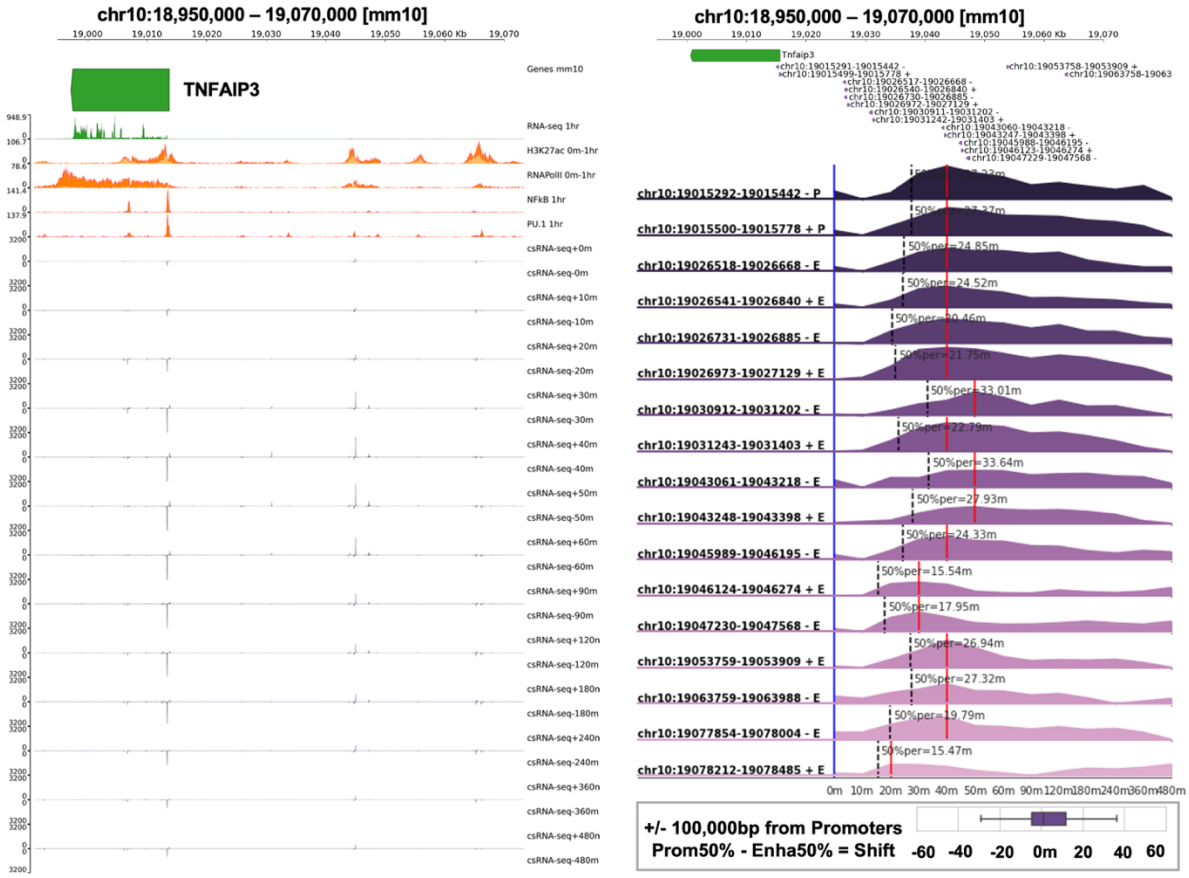

Figure 3.7 Genomic overview of the TNFAIP3 locus along with 50th percentile calculations and overall shift of all enhancer promoter pairs genome-wide

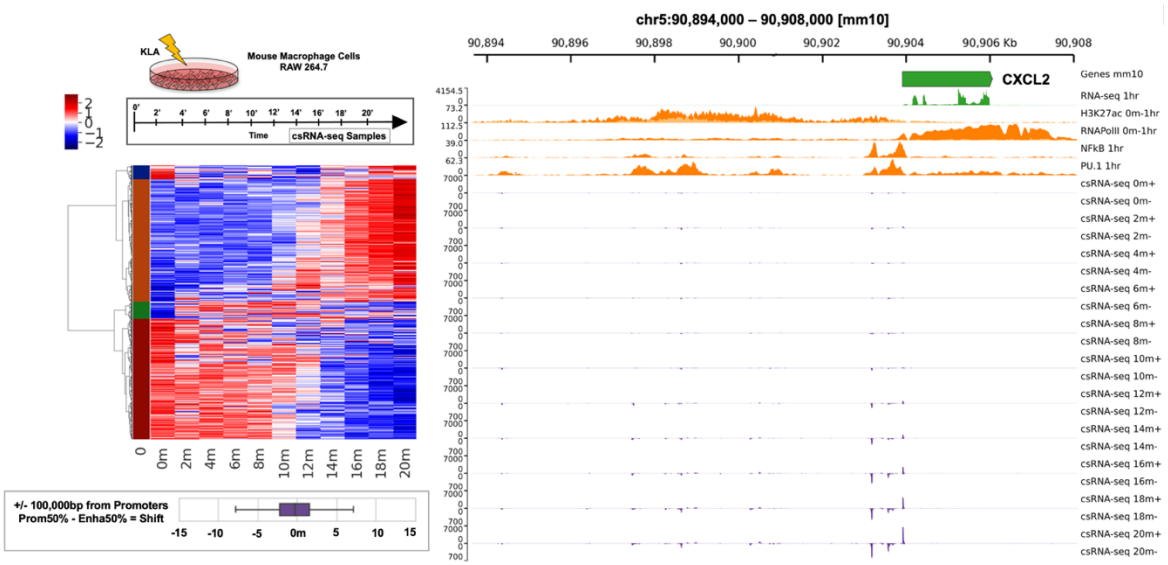

Figure 3.8 Shift calculation for the short interval csRNA-seq time series dataset.

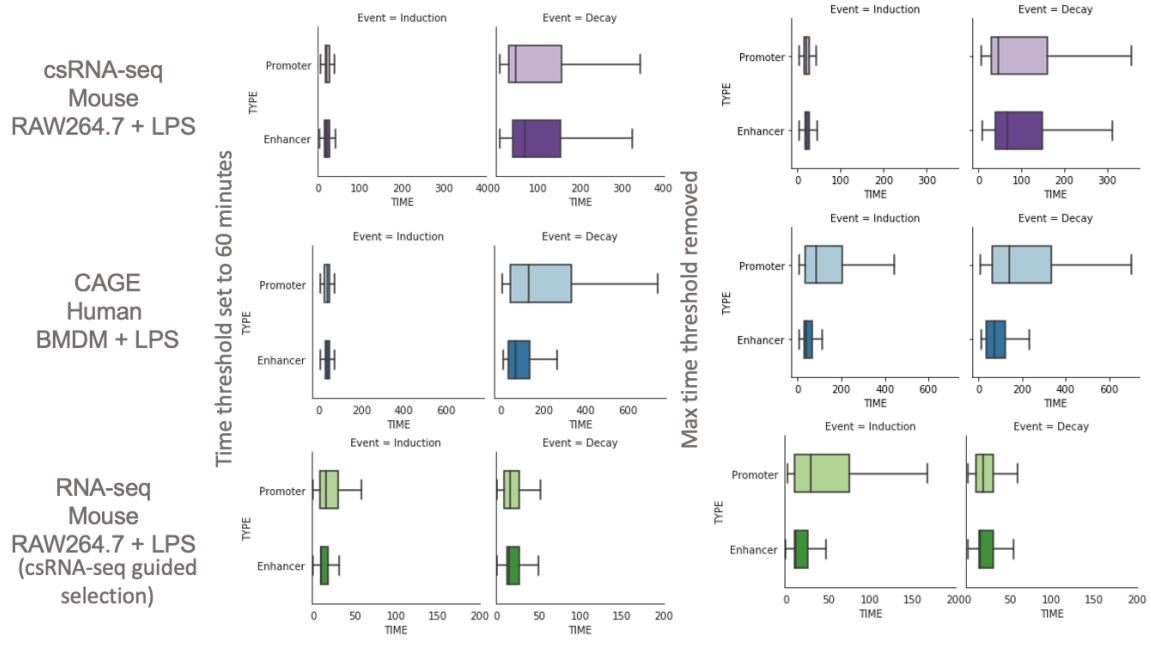

Figure 3.9 csRNA-seq, CAGE, and RNA-seq induction and decay calculation results. Results shown represent limiting induction max time to 60mins and allowing no maximum time point.

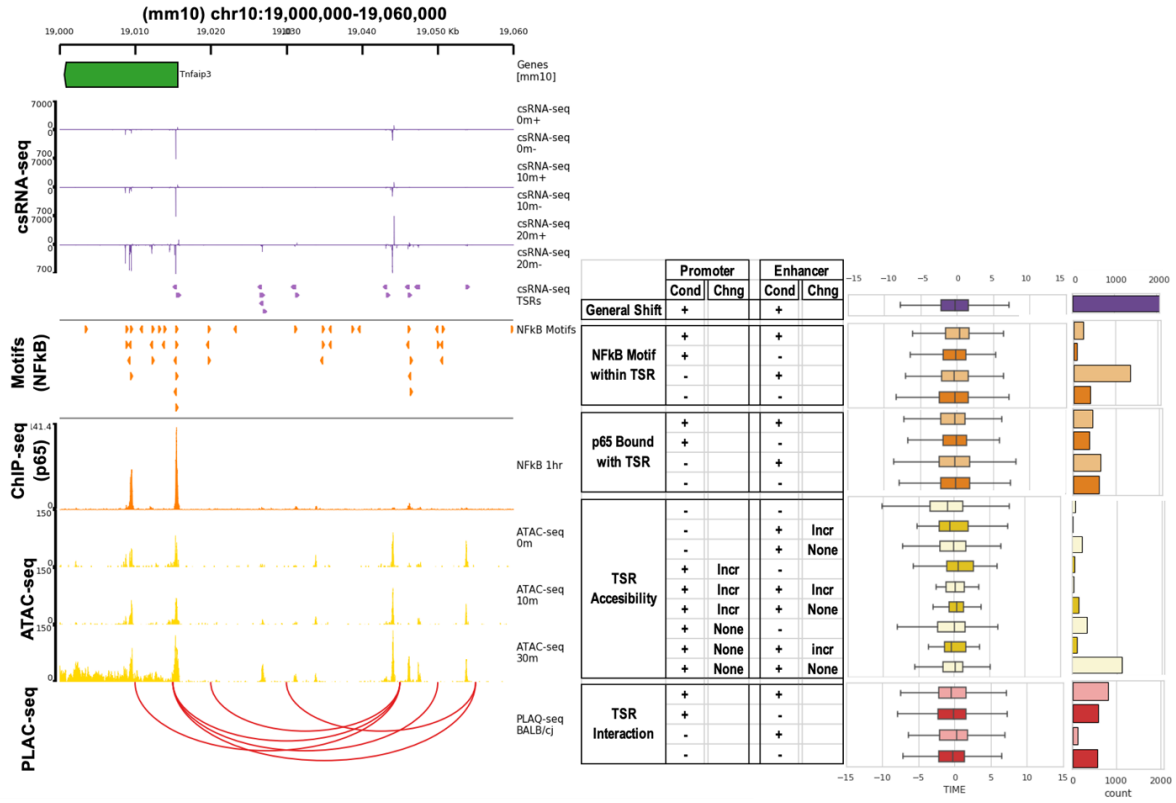

Figure 3.10 Shift calculation results after annotating enhancers and promoters based on various genomic attributes.

Acknowledgements

Chapter 3, as a whole, has been compiled as pending publication. Hillary, R., Heinz, S., Benner, C. The dissertation author will be the primary investigator and author of this paper.

Chapter 4 Characterization of Localized Induction: Cis-Regulatory Domains

4.1 Abstract

The mechanisms governing potential regulatory interactions between regulatory elements located proximal (promoters) and distal (enhancers) to gene bodies remains elusive. A current model of enhancer/promoter mechanisms suggests that physical contact between enhancers and promoters mediates gene expression (Bulger & Groudine, 1999). However, our results have

identified individual genomic locales containing many regulatory elements whose transcriptional activities are highly correlated and induced simultaneously and with similar kinetics. This localization of correlated induction suggests the formation of regulatory domains of transcription activity whose functional contribution to gene expression will be interesting to study in the future. Here we report the development of a tool to identify these genomic regions of correlated transcription activity, referred to as cis-regulatory domains (CRDs) from time series datasets, specifically csRNA-seq. We characterize the prevalence of CRDs genome wide and identify them as potential key loci in the modulation of the TLR4 transcriptional response.

4.2 Methods

4.2.1 Computationally Define Cis-Regulatory Domains

To computationally define CRDs we will leverage csRNA-seq time series data. Each individual datapoint, assembled according to linear time, yields individual datapoints we will refer to as transcription activity profiles. These profiles represent individual TSS activity (**Figure 4.1A**) that can be leveraged to directly compare neighboring TSS transcription activity patterns by calculating their Pearson correlation coefficient. This pairwise comparison is applied to the entire chromosome of TSS yielding an "all vs all" matrix of individual Pearson correlation coefficients for each possible TSS to TSS transcription activity profile comparison. (**Figure 4.1B**)

Using this matrix of coefficients (R) as the main data structure we then begin to identify groups of adjacent TSS (i.e., CRDs) that are highly correlated with one another. Specifically, starting from a given TSS as the primary member of a putative CRD, our method test if CRD boundaries should be expanded by considering the inclusion of adjacent upstream and downstream TSS, calculating the mean TSS Pearson (R) of the group as they are added. (**Figure 4.1C**) The threshold of the mean group must be greater than an empirically determined arbitrary value (here:

0.4) to continue. In conjunction with the mean correlation within the group it examines the individual candidate upstream and downstream TSS Pearson (R) individually to assess if their Pearson (R) Coeff is greater than the threshold of 0.4. If these thresholds are met the upstream and downstream TSS are included in the CRD. If the upstream or downstream TSS correlation is below the threshold a penalty is given for that upstream or downstream TSS, two penalties are allowed for the current application and guide the identification of CRD boundaries. The upstream and downstream TSS will be assessed independently until both directions accrue enough penalties or if the mean correlation of the groups of TSS falls below the specified threshold. (**Figure 4.1D**) When these criteria are met, the boundaries for the CRD are stored and the process begins anew for the next downstream TSS to become the focus TSS. (**Figure 4.1E**)

Groups of correlated TSS, are stored during this process and CRD candidates are identified by identifying multiple overlapping boundaries. In short, each TSS will produce candidate boundaries, overlapping candidate boundaries form candidate CRDs. These candidate CRDs are then further assessed to assert that the mean matrix generated by these overlapping candidate boundaries satisfy a mean matrix Pearson (R) of 0.4 or greater as well as a size threshold containing at least 10 TSS. Those groups of boundaries that satisfy these criteria are then identified as the final CRDs. Groups of boundaries that do not satisfy the criteria are also stored for algorithmic optimization and testing purposes.

4.2.2 Timeframe Rational for CRD Selection Criteria

Transcription and its regulation within robust immune responses is highly dynamic. As discussed within chapter 2 we observe that key canonical transcription factor motifs, including NF-κB, AP-1, and IRF, each display enrichment within hierarchically clustered TSS. This provided evidence for a transcriptional cascade of regulatory signals within distinct time intervals.

For our study we will examine CRDs forming during the earliest stages of the TLR4 pathway induction, specifically to TSS induced within the first 30 minutes of LPS exposure. The rationale for selecting 30 mins is due to the transcription factor NF-κB being known to enter the nucleus within 10 mins of LPS exposure, along with our defined induction interval in which we see NF-κB motif enrichment within our time course (Ding et al., 1998). Within this interval we capture a group of CRDs that represent the principally induced regions within our model systems and represent a sizable sample to carefully deconvolute function and purposes of CRDs as well as address further questions such as potential enhancer/promoter communication within such domains.

4.3 Results

4.3.1 Identification of Early Interval Cis-Regulatory Domains

Using our CRD algorithm described above, a total of 153 CRDs were identified within our short interval time series. (**Figure 4.2**) These identified CRDs encompass the early response loci within the TLR4 immune pathway. Of these CRDs the mean size was 63,763 bp with the largest CRD being 217,987 bp in size. The mean distance of TSS from one another within these CRDs is 8012 bp and the mean correlation of the CRDs is 0.47. Each CRD can also be assigned a transcript activity type giving their mean activity profiles, if a CRD is increasing in transcription activity it is considered induced, while if the mean activity profile is showing a reduction in activity or no change in activity they are identified and labeled accordingly. Within our 153 CRDs the majority are induced [N= 98], while some are showing a reduction in transcription [N=47] and the remaining few generally show little to no change in transcription over time [N=8].

When examining CRDs we observe canonical motifs enriched at induced CRD TSS, namely AP-1 motifs along with NF-κB. The reduced CRDs are principally enriched for motifs

recognized by various ETS factors. Using the downregulated CRD TSS as backgrounds and examining the induced CRD TSS as the foreground we found the top, most significant enriched motifs to be NF-κB -p65, NF-κB -p65-Rel and NF-κB -p50, p52, indicating that this enrichment of NF-κB-containing sites represents we are indeed examining the initial TLR4 response. (**Table 4.1, Table 4.2 & Table 4.3**) Furthermore, we performed gene ontology (GO) analysis using GREAT (McLean et al., 2010) to examine the TSS within the induced CRDs and results confirmed the genes associated with these TSS are highly enriched for immunological pathways and cellular responses to biotic stimuli, including specifically genes responsible for cellular responses to lipopolysaccharide (LPS). (**Figure 4.3**)

4.3.2 CRDs and Super Enhancers

After characterizing CRDs called from csRNA-seq from macrophages exposed to LPS within 30 mins we decided to directly compare these loci with other identifiable regulatory entities, specifically super-enhancers. Super-enhancers represent a loci containing multiple, putative enhancers that are defined by their unusually high transcription factor binding activity through ChIP-seq (Pott & Lieb, 2015). With the identification of CRDs there is a need to justify if they are an entity separate or perhaps identical to other known, even if loosely defined, regulatory bodies. Within our study CRDs are the called using correlated time series csRNA-seq data. In contrast, super-enhancers are typically identified by detecting highly enriched regions of ChIP-seq signal, specifically H3K27ac. We used ROSE to detect super-enhancers within an H3K27ac ChIP-seq dataset generated from LPS-treated macrophages and harvested after 30 minutes of exposure (Whyte et al., 2013, Lovén et al., 2013). From this analysis we detected 430 super-enhancers, of which only 84 super-enhancers (SE)overlapped with CRDs. (**Figure 4.4**) Of these that overlapped with CRDs only 72 CRDs reported an overlap with one or more SEs, resulting in the larger

proportion of CRDs [N=80] not overlapping with known SEs. This result suggests SEs can be a subcomponent of a CRD but SEs aren't a crucial subcomponent for CRD function or formation. (**Figure 4.5, Figure 4.6**) Furthermore, this gives evidence that CRDs are not identical entities with SEs but its own genomic phenomena.

4.3.3 Reconciliation of CRDs with 3D Genome Structure

To further characterize CRDs we performed a direct comparison with 3D spatial chromatin mapping data to establish if 3D interactions are required for CRD formation and function. To reconcile the formation with CRDs with known genomic 3D structures we examined known characteristics associated with topological associated domains (TADs). (**Figure 4.7**) Using a suite of tools available through cooltools (Nezar Abdennur et al., 2022) and HiCExplorer (Wolff et al., 2020) we examined Hi-C data representing the genomic 3D compartmentalization of LPS-treated macrophages at 0 m and 60 m after stimulation. These samples were both sequenced deeply enough to ensure interpretation on finer resolutions is appropriate, for these analyses we examined Hi-C interactions down to a 10-kilobase binned resolution. (**Figure 4.8**) First, we examined if there are higher levels of 3D interaction within CRD boundaries, suggesting CRDs function is associated with proximity. As a means of direct comparison of our CRD vs a randomized background we generated 1000 sets of 153 CRDs whose bounds were specifically size selected to match the boundaries within the CRD pool. When examining the interaction values of CRD vs randomized background we detected a relative overall increase of 3D interactions of both groups at 60 min after LPS exposure compared to untreated cells. We then directly compared CRD interaction values with the randomized background using either the mean Hi-C interaction within the boundaries or the aggregate Hi-C interaction values. Both approaches revealed higher Hi-C interactions within the CRD pool relative to background. (**Figure 4.9**)

This suggests that 3D interactions are enriched in CRDs and potentially reveal 3D interactions as an important mechanism driving CRD formation and function. We also pursued an approach that binned Hi-C interactions based on distance within CRDs and randomized background Hi-C interaction comparisons which again yielded the same result that CRDs are indeed more highly enriched for 3D genome interactions within their boundaries relative to background. (**Figure 4.9**)

Leveraging known TAD characteristics, we examined if CRDs share similar or dissimilar characteristics in order to identify if CRDs are their own genomic entity and how they relate to other known genomic structures. Calling TADs macrophages exposed to LPS for 60 mins resulted in identification of 8538 TADs, due to the finer 10-k binning resolution used we acknowledge many of these TADs could be considered "sub" TADs of a larger TAD structure. We also called TADs at 100kilobase resolution, yielding 1702 TADs. Our CRD dataset is considered sparse due to it being defined only by detectable TSS using csRNA-seq, and not all regions of the genome are active, captured or otherwise represented. In contrast to our CRD dataset the Hi-C data is more data rich in that basically each individual 10-kb region of the genome contains signal with little to no empty bins. Given these data properties between our CRD and TAD datasets we attempted a best-practice "like for like" comparison between these two complex yet loosely defined genomic structures. Using TADs called at 10 kb resolution we detected that the majority of our CRDs are found within TADs. At this resolution we detected 130 CRDs that are fully contained within TADs, with 25 CRDs appearing within a TAD but also extending outside one of the two boundaries. CRDs do not appear to span the entirety of TAD boundaries. Using three categories: In-TAD, Exceeds-TAD and Not-in-TAD, along with size-selected randomized genomic intervals, we were able to generate expected counts for each of these categories. Examining 10 kb-resolution

TADs and our expected counts for the categories described we found that our observed CRD counts for these categories to be significant, in that we found that the frequency that CRDs were found within TADs relative to background to be higher than random. While examining the amount of 3D genomic loops formed within our CRDs compared to randomized intervals we found after 60m of LPS exposure CRDs appeared to be enriched for these loops [N=125 Loops] relative to loops found within our randomized dataset [μ = 75 Loops]. These findings suggests that these CRDs are a subcomponent of TADs.

To assess other known TAD characteristics in comparison to CRD boundaries we then examined enrichment of genomic attributes at their boundaries as well as within their boundaries, notably enrichment of CTCF among other factors (Nanni, Ceri, & Logie, 2020). For each analysis the data is binned into 5 kb bins, while the CRD or TAD body is represented in bins of length percentages (10%, 20% etc) to allow a more equalized comparison between the differing boundary sizes. (**Figure 4.10**) TADs demonstrate their characteristic CTCF enrichment at its boundaries while CRDs exhibit CTCF enrichment at boundaries to a lesser degree. (**Figure 4.11B**) Using cooltools we leveraged their insulation scoring (Nezar Abdennur et al., 2022), where low values indicate potential boundaries and/or decreases in 3D interactions and high values represent rich 3D genomic interactions. (**Figure 4.11A**) Using this insulation score metric we observe the expected depletion of 3D interactions at TAD boundaries. For CRDs there is a decrease in insulation at CRD boundaries, but overall, the insulation scores are much higher within CRDs. This suggests that while insulation scores are the highest within CRD bodies and have a definitive decrease in insulation scores at boundaries, the overall elevation in signal suggests they are already within a locus rich with 3D interactions. (**Figure 4.11A**) We applied the same approach using H3K27ac, p65 and PU.1. We examined PU.1 in this context due to its critical transcription

regulator roles during immune cell development (Li, Hao, & Hu, 2020) and found that, in contrast to TADs, CRD bodies are enriched for these TF and histone modifications relative to regions outside of CRD boundaries. This suggests that CRDs are not only 3D interaction-rich environments but also exhibit highly elevated TF binding within their boundaries. (**Figure 4.11B**) This additional information suggests that not only are CRDs subcomponents of TADs but specifically of transcriptionally active sub-TAD modules.

4.4 Discussion

The formation, identification, and characterization of LPS-specific cis-regulatory domains represents a suitable genomic entity to better understand focused and intense regulatory dynamics at play. Here we describe how CRDs can be computationally defined and how their genomic attributes identify them as genomic entities independent of other known genomic entities such as TADs and super-enhancers. We establish where CRDs fit into the context of known genomic constructs and characterize their unique attributes. CRDs appear to be transcriptionally active sub-TAD domains and can, but not always, contain super-enhancers.

CRDs showing strong induction of TSS within their boundaries are enriched for NF-κB sites relative to downregulated domains. This provides evidence that these specific sites could be driving recruitment and transcription initiation for all TSS within our defined CRD. The identified TSS within CRDs represent many genomic loci to target for perturbation in future experiments. If not all TSS have NF-κB binding sites but are induced, then by what means are they being activated transcriptionally? We again assert that transcription among these TSS within these domains could be the result of communication between target regulatory elements and beneficiary TSS. An emerging challenge in the field of enhancer biology is to address what functional role enhancer RNAs (eRNA) contribute transcription regulation. eRNAs are unstable, targeted for degradation

and not translated like messenger RNAs, but their function is unknown. (Zhang et al., 2019) This work has identified that induced TSS that represent a pool of candidate eRNAs whose function could be further elucidated in future studies. If eRNAs are not important direct targets for transcription regulation, perhaps they contribute to regulation and modulation of gene expression within CRDs though other mechanisms.

Our results show a significant enrichment for 3D interactions within CRDs relative to size-selected and randomized genomic intervals. This result suggests that CRD function is associated and perhaps dependent on 3D genome interactions not only for their phenotype but also their function. This observation is in line with liquid-liquid phase separation (Sabari et al., 2018).

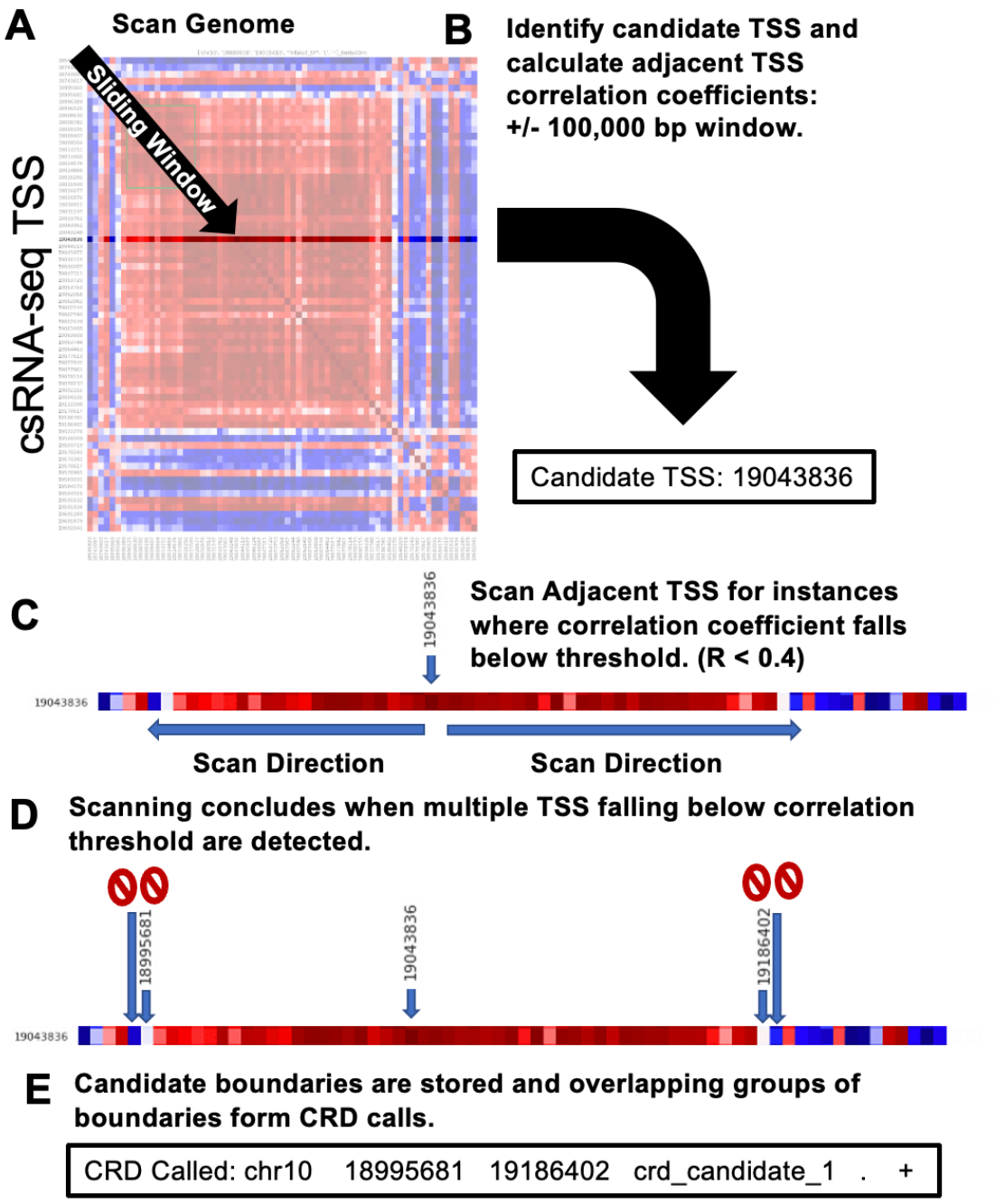

Figure 4.1 Schematic of a CRD detection algorithm. A) A sliding window approach is used. B) Each TSS is examined independently. C) The algorithm scans bidirectionally, testing all adjacent TSS. D) Stop case for algorithm, detection of boundaries. E) Algorithm output.

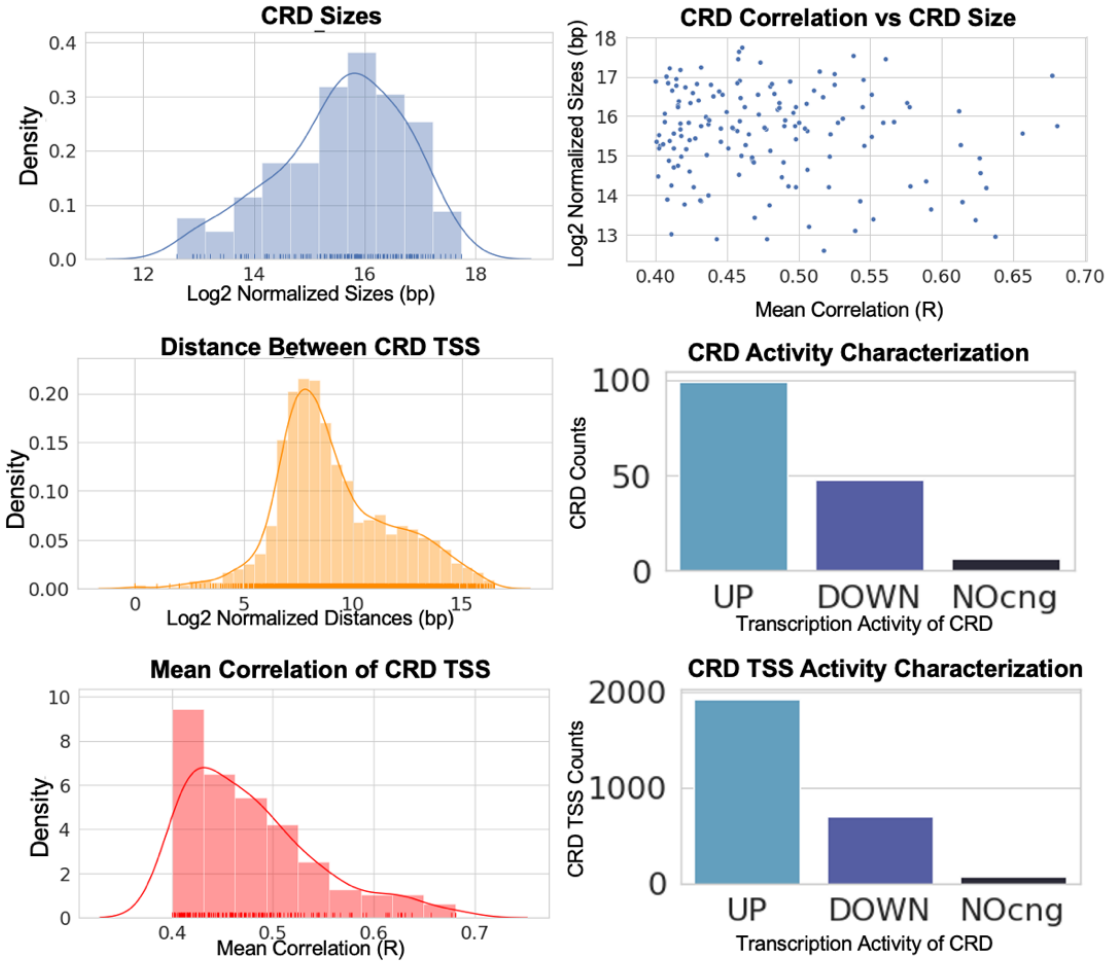

Figure 4.2 Summary statistics of 153 CRDs detected within macrophages after 30 minutes of LPS exposure.

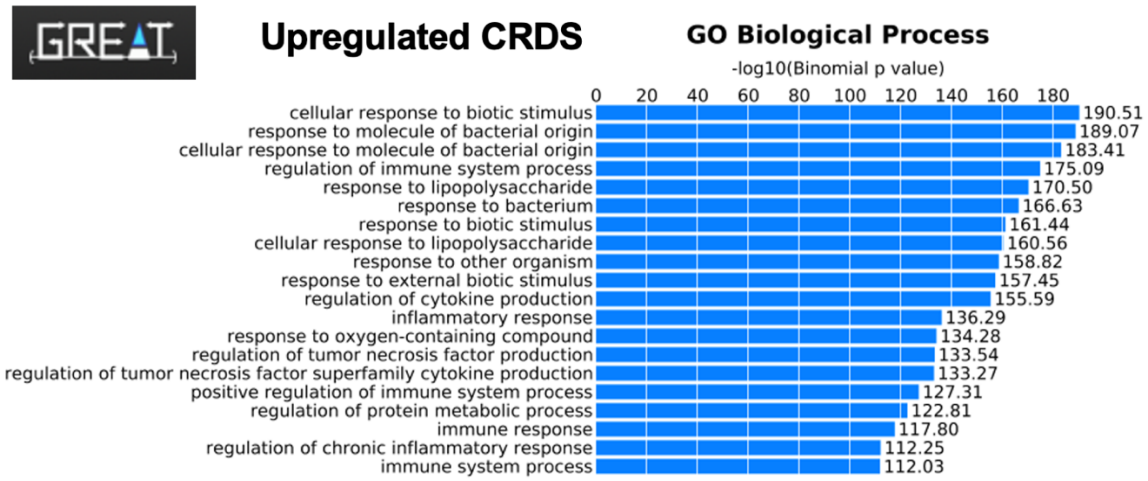

Figure 4.3 Gene Ontology (GO) results.

Table 4.1 Motif enrichment of induced CRD TSS with the repressed CRD TSS used as background TSS. Induced CRD TSS have strong enrichment of NF-κB motifs along with an enrichment of CEBP containing regulatory elements.

Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif
1		NFkB-p65(RHD)/GM12787-p65-ChIP-Seq(GSE19485)/Homer	1e-78	-1.807e+02	0.0000	376.0	19.49%	46.5	6.64%
2		CEBP:AP1(bZIP)/ThioMac-CEBPb-ChIP-Seq(GSE21512)/Homer	1e-38	-8.785e+01	0.0000	319.0	16.54%	53.0	7.58%
3		NFkB-p65-Rel(RHD)/ThioMac-LPS-Expression(GSE23622)/Homer	1e-35	-8.176e+01	0.0000	96.0	4.98%	7.1	1.02%
4		NFkB-p50,p52(RHD)/Monocyte-p50-ChIP-Chip(Schreiber_et_al.)/Homer	1e-35	-8.176e+01	0.0000	96.0	4.98%	7.2	1.03%

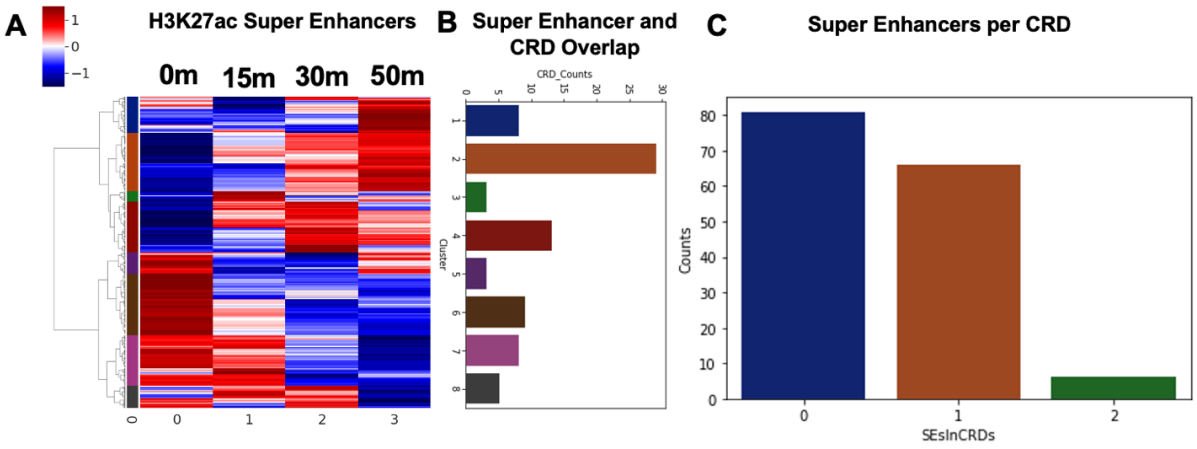

Figure 4.4 Super enhancer analysis results. A) Heatmap of super-enhancers detected at individual timepoints. B) Super-enhancers and their respective cluster from panel A and that overlap with CRDs. C) Counts of how many super enhancers were detected within CRDs.

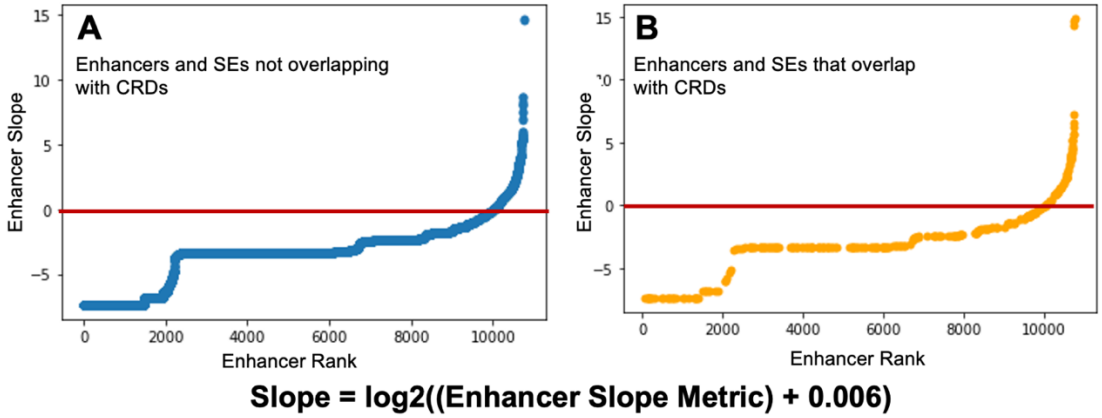

Figure 4.5 Enhancer Slope vs Enhancer Rank graphs. Plots are separated based on enhancers/super enhancers overlapping with CRDs or not.

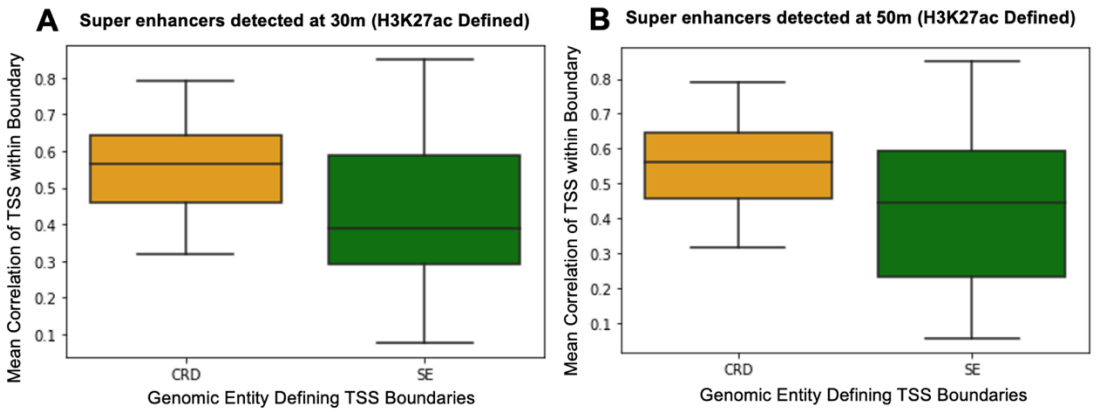

Figure 4.6 Plots of mean TSS correlation values comparing CRD TSS and TSS within super enhancers. Plots represent super enhancers detected at 30 minutes (left) and 50 minutes (right).

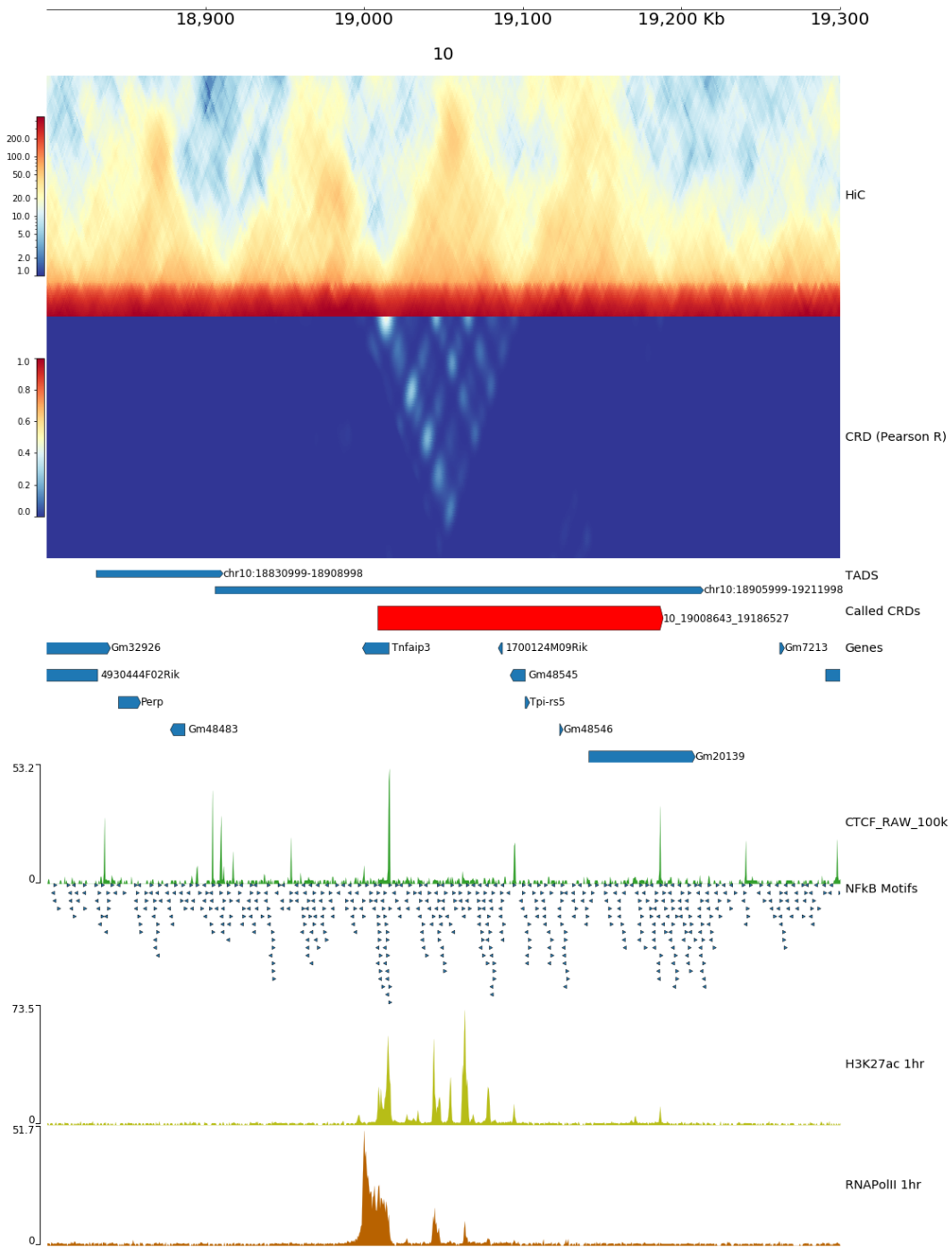

Figure 4.7 Overview of Hi-C data and a CRD plotted in the same genomic space. Genes, CTCF binding, NF-κB motifs, H3K27ac activity and RNA Polymerase II binding within the region are displayed.

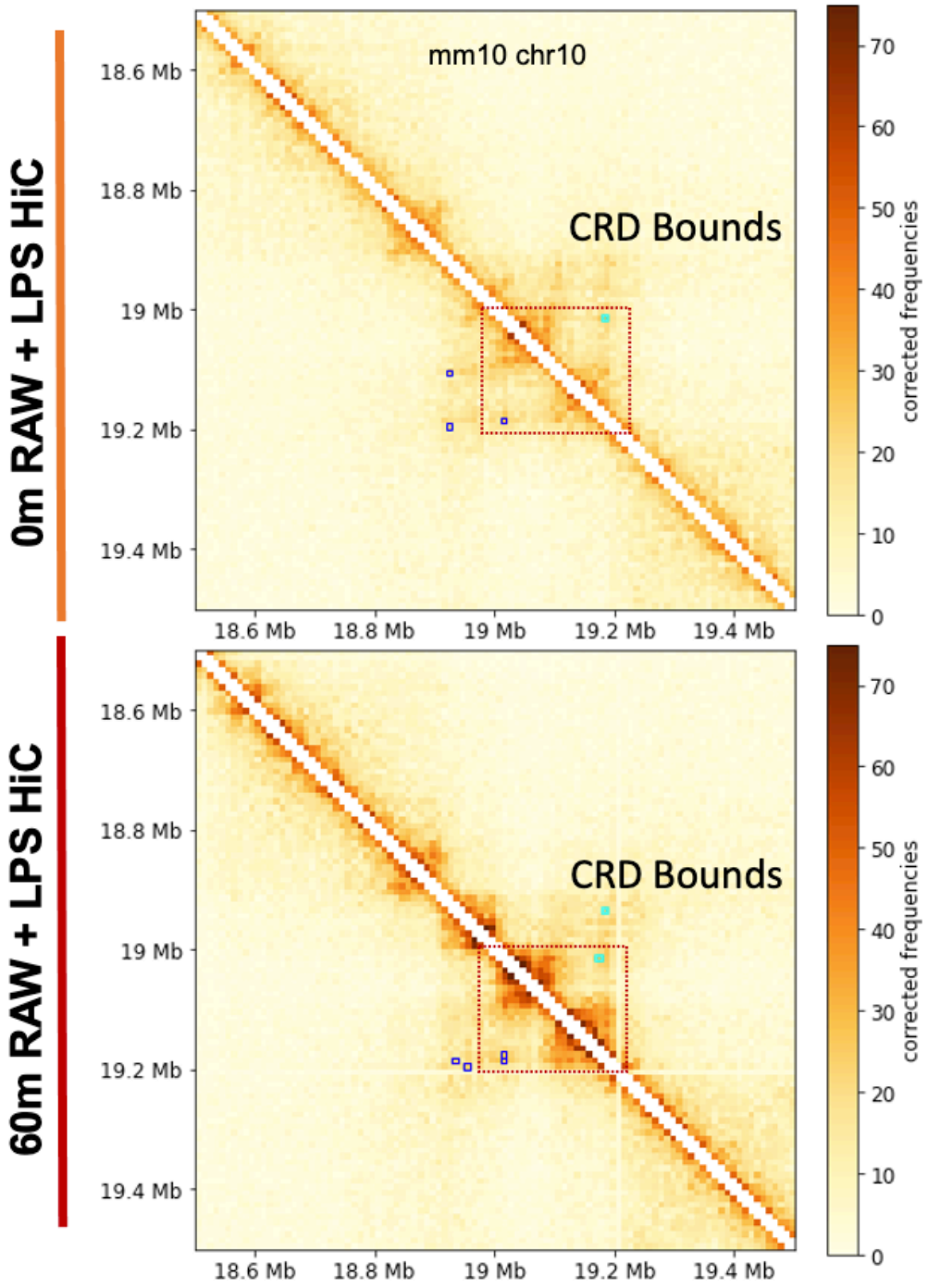

Figure 4.8 Hi-C showing genomic 3D interactions at 0m and 60m of LPS exposure within RAW 264.7 cells. A single CRD boundary is shown.

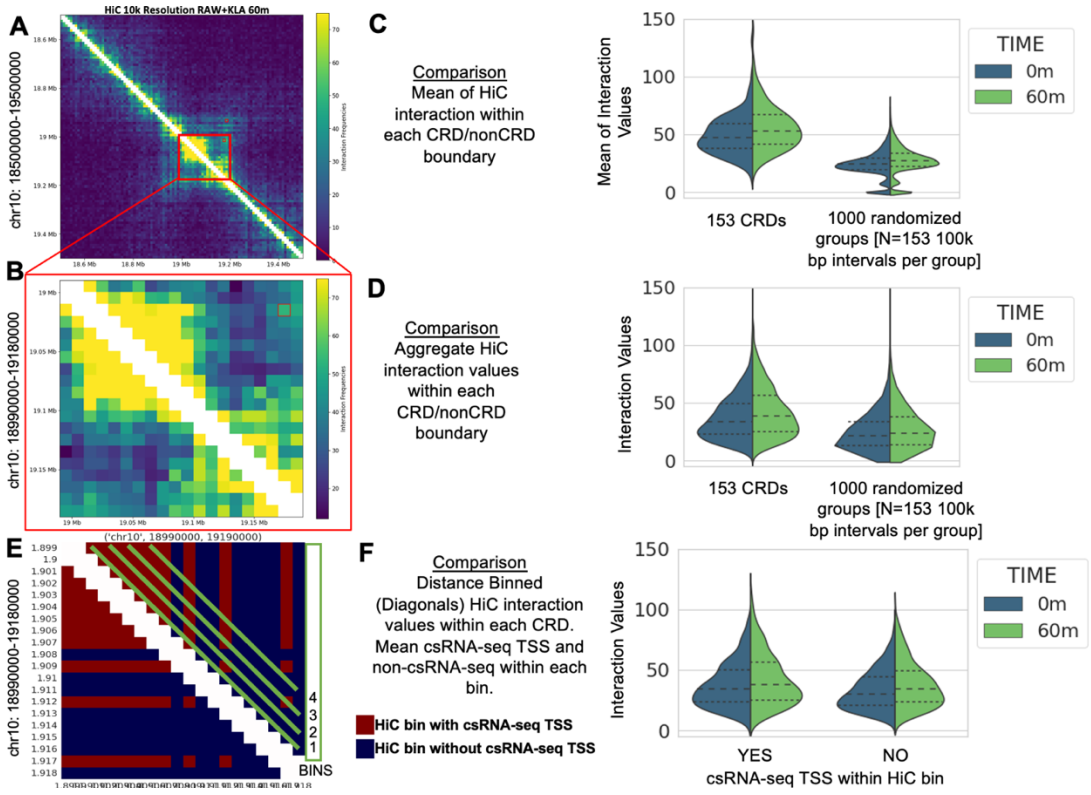

Figure 4.9 3D genome interaction analysis results examining interactions within and outside of CRDs.

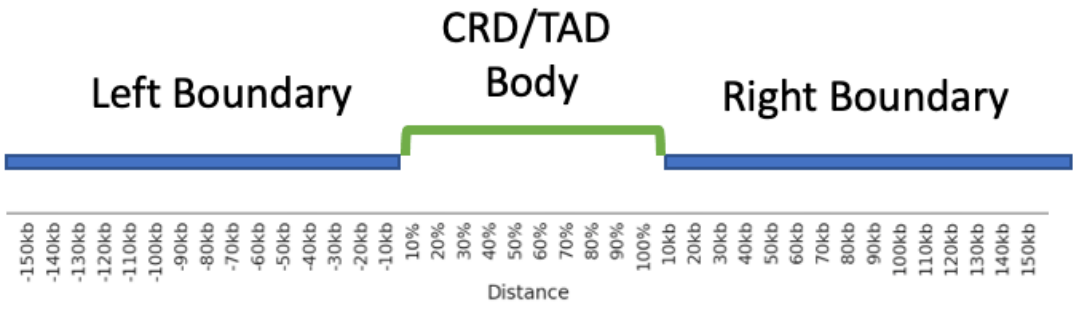

Figure 4.10 Schematic showing the data collection criteria for directly comparing CRDs and TADs. Schematic displays how the data in Figure 4.11 is organized.

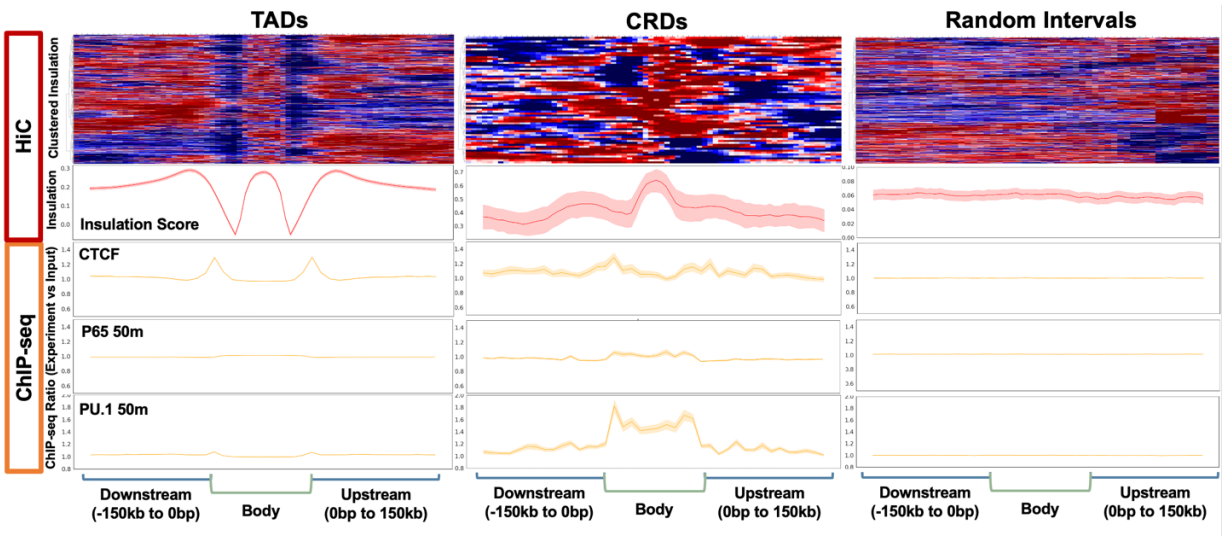

Figure 4.11 Results directly comparing TADs (left), CRDs (center) and random genomic intervals (right) showing the insulation scores, and ChIP-seq detection (CTCF, p65 and PU.1) for each genomic attribute.

Acknowledgements

Chapters 1,2,4,5,6 represent research intended to be published together pending further insight provided by a Dual-MPRA approach described in Chapter 6. Hillary, R., Guzman, C., Heinz, S., Benner, C. The dissertation author will be the primary investigator and author of this paper.

Chapter 5 Degradation of Genome 3D Structure and Leveraging Natural Genetic Variation

5.1 Abstract

Using csRNA-seq to examine robust immune responses within macrophages revealed that transcription induction between localized enhancers and promoters is near simultaneous, their transcription activity kinetics are highly correlated with adjacent regulatory elements, and 3D interaction at these transcriptionally active loci is significantly elevated. Given these observations along with general uncertainty if and how these regulatory elements communicate, it would be prudent to examine if these regulatory elements maintain similar regulatory dynamics if specific

attributes of the system are perturbed. Specifically, we examined CRD activity when 3D chromatin structure is perturbed as well as examined if natural genetic variation affects transcription at individual TSS and adjacent TSS, suggesting localized transcription activity is the result of communication. In cohesion-depleted and poly(I:C)-treated HCT116 cells correlation of TSS activity within CRDs defined in cohesion-expressing cells was significantly diminished, suggesting 3D interaction is a critical component of correlated TSS transcriptional activity. Leveraging the natural genetic variation (NGV) between bone marrow-derived macrophages from two different mouse strains (C57BL/6Bl/6 and SPRET) exposed to LPS we find not only do mutations in NF-κB sites show diminished transcription activity but also their adjacent non-mutated sites. Together, these results give key insight into CRD function and evidence of regulatory element communication.

5.2 Methods

5.2.1 CRDs in other cell types, species, and stimuli

We examined the prevalence of CRDs outside of our model system using our novel CRD algorithm. For this approach we leveraged HCT116 colorectal cancer cells exposed to polyinosine-polycytidylic acid (poly(I:C)), a toll-like receptor 3 (TLR3) immune pathway agonist that produces a regulatory cascade similar to KLA (TLR4). Poly(I:C) was used for these experiments because HCT116 cells do not respond to KLA. (**Figure 5.1**) We used this model system not only to examine CRD formation outside of our macrophage model system but specifically due to the RAD21 degron system engineered in these cells that permits the auxin-dependent depletion of cohesin, which is necessary for TAD formation. (Rao et al., 2017) This permits direct examination of CRD formation in the presence of cohesion as well as the absence of this crucial genome 3D structure component. Cells were harvested at timepoints 0m, 15m, 30m, 60m and 120m after exposure to

poly(I:C), in both control and auxin-treated conditions (which results in cohesin depletion). Alignment, analysis and QC was performed using HOMER, (Lin et al., 2012) and only TSS whose expression levels replicated between the two replicates were considered. (**Figure 5.2**)

5.2.2 C57BL/6 and SPRET BMDM Natural Genetic Variation Curation

To examine localized transcription activity and to assess if natural genetic variation effecting specific transcription factor binding motifs impacts regulation, we analyzed transcription similarities and differences between two mouse strains, C57BL/6Bl and SPRET. These two mouse strains were selected to leverage the unique genetic properties of SPRET mice whose genome containing ~40 million sites of natural genetic variation compared to C57BL/6 (Keane et al., 2011). We specifically examined bone marrow derived macrophage cells between these two strains. These cells were treated with KLA for 1h and multiple capture methods were used to examine transcription and epigenetic effects. (**Figure 5.3**) (csRNA-seq, ChIP-seq, ATAC-seq) ChIP-seq and ATAC-seq raw data were obtained from Link et al. (Link et al., 2018). The csRNA-seq data was generated in-house. All datasets were aligned and processed following the workflows described within HOMER (Heinz et al., 2010). To leverage the natural genetic variation each TSR found within the SPRET and C57BL/6 csRNA-seq datasets were centered on their TSS.

C57BL/6 was selected due to its similarity to our initial model system using RAW264.7 macrophage cells, SPRET was selected due to its substantial amount of genetic variation relative to our model system and C57BL/6. We analyzed TSS that arise from homologous DNA found in both strains so that natural genetic variations, such as SNPs and INDELS, found within both strains could be fairly compared. Given natural genetic variation can affect sites of known transcription factor binding motifs we scanned for instances of TLR4 canonical motifs within -200bp/+50 bp of our TSS and assessed if known SPRET genetic variants could either enhance or diminished

binding of transcription factors to these sites. Instances of these motifs occurring within -200bp/+50bp of motifs of interest were scored against the consensus motif in question. Sites showing a genetic change that resulted in a more optimal motif score for SPRET was given a positive score, sites where SPRET genetic variation negatively affected the motif sequence resulting in a poor consensus match resulted in a negative score. While the positive/negative effects of natural genetic variation were specific to SPRET, the inverse of the scoring can be representative of motif binding affinity for C57BL/6 relative to SPRET. These steps of assessing changes to transcription factor motif binding affinity were done using HOMER annotatePeaks. (Heinz et al., 2010)

5.3 Results

5.3.1 Loss of 3D structure impacts CRD form and function.

After applying our CRD calling algorithm to the HCT116 time series datasets we detected 39 individual CRDs in samples with cohesin intact and 30 CRDs in samples with cohesin depleted with 26 CRDs shared between the two conditions. Within our HCT116 time series CRD pools we replicated our TAD vs CRD results in that CTCF, H3K27ac and insulation score activity patterns were highly similar. (**Figure 5.4**) While the CTCF and insulation score attributes and plot shapes were similar, we found an overall depletion of these values within the cohesion-depleted CRD pool boundary analyses, suggesting a reduction in 3D structural integrity. When CRDs are examined between the two pools, specifically looking at the mean Pearson (R) correlation of TSS profiles contained within CRD boundaries between both datasets, we detected that overall, the mean correlation was much higher in cohesin-intact time series data CRDs in comparison to the cohesion-depleted dataset. (**Figure 5.5**) The higher mean correlation was also observed when TSS between the two time series approaches were matched such that only TSS expressed in both

datasets were considered. This suggests the depletion of cohesion, resulting in the degradation of 3D genome structure, appears to disrupt the correlation between TSS transcription activity profiles within CRD boundaries. (**Figure 5.6**) This loss of 3D structure along with the loss of highly correlated transcription activity profiles within CRD TSS provides evidence that 3D structure is a crucial component of CRD formation. This loss of regulatory element 3D interaction could also be evidence of a loss of influence or communication at these individual TSS. (**Figure 5.7**)

5.3.2 Effects of Natural Genetic Variation on TSS and Adjacent TSS Transcription

To collect evidence of inter-regulatory element communication we leveraged the natural genetic variation between two mouse strains: C57BL/6 and SPRET. To identify candidate TSS of interest for direct SPRET and C57BL/6 BMDM TSS induction after 60 min LPS exposure we focused on peaks whose induction over no treatment was 2-fold or higher. For the two mouse strains this selection criterion resulted in two pools of 17208 csRNA-seq-identified TSS within the C57BL/6 dataset and 16166 csRNA-seq-identified TSS for the SPRET dataset relative to SPRET and C57BL/6 cells not treated with LPS. (**Figure 5.8**) To compare the two strains, we took the induced sites of one strain and examined their induction activity within the other strain. Furthermore, among the candidate sequences to be examined, we annotated their status by the presence of a TLR4 signaling-associated motif and if a mutation, given the natural genetic variation, might overlap the motif to diminish or enhance the binding of the corresponding transcription factor. The motifs examined included in this approach were the following: AP-1, CEBP, CEBP: AP1, CRE, CTCF, E-box, ETS, GFX, GFY, ISRE, MITF, NFY, NF-κB-65, PU.1, RUNX-AML, STAT1, Sp1, and YY1. (**Figure 5.9 Figure 5.10 Figure 5.11 Figure 5.12**)

First, we examined induced TSS and their adjacent TSS within +/-500 bp of the candidate TSS which will include the TSS in question as well as TSS located proximal to a mutated TSS

site. For both the SPRET and C57BL/6 induced candidate TSS pools we identified that of those TSS with predicted diminished binding for mutated AP-1, CEBP, NF-κB-p65 and PU.1 sites due to natural genetic variation to have lower transcription activity relative to their non-mutated counterparts. Extending our window of analysis to examine only TSS within +/-500bp-5000bp of TSS containing motifs of interest we see this trend continue but most significantly for AP-1 and NF-κB-p65 containing sites. (**Figure 5.9 Figure 5.10 Figure 5.11 Figure 5.12**)

We then focused specifically on predicted functional and dysfunctional NF-κB sites within our C57BL/6 and SPRET mouse strain datasets. Dysfunctional in this context meaning the natural genetic variation occurring within SPRET relative to C57 represents a diminished motif match to the consensus sequence. We examined TSS located within 0-500bp, 500-5000bp described previously, but we extended our analysis further to include TSS located within 5001-20000bp, 20001-50000bp and 50001-100000bp. Specifically for NF-κB we found that significant differences between functional and dysfunctional NF-κB TSS induction extended as far out as 20,000bp away from candidate TSS. (**Figure 5.13**)

We further annotated mutated SPRET NF-κB sites to group candidate TSS based on two criteria. First, if the candidate TSS had a mutated or wild-type NF-κB site and second if adjacent TSS contained mutated or wild-type NF-κB TSS within +/- 50000bp. This resulted in 4 groups of candidate TSS in which induction patterns can be directly compared. The group of TSS where the candidate TSS was a mutated NF-κB who's adjacent TSS had no NF-κB sites nearby showed the least induction at the TSS as wells as adjacent TSS out to 50000 bp. We identified that if the NF-κB site was mutated but within the 50000bp window a normal NF-κB site existed the mean TSS induction was similar to non-mutated NF-κB candidate and adjacent TSS activity. (**Figure 5.14**)

5.3.3 Saturation of NF-κB sites within Induced CRDs

We examined within our CRD the number of sites that contained NF-κB-like motifs within their boundaries. We coupled this information with an assessment of if the TSS within the CRD are all being induced or repressed within macrophages exposed to LPS. We found that CRDs induction state is correlated with a higher saturation of NF-κB transcription factor binding motif containing TSS relative to repressed CRD TSS. (**Figure 5.15**)

5.4 Discussion

In this study, we demonstrated that the loss of 3D genome structure and natural genetic variation affect CRD formation, TSS induction potential, and adjacent TSS transcription activity. By leveraging the RAD21 degron system deployable within HCT116 cells (Rao et al., 2017) we observe the degradation of correlated TSS activity within CRDs. This observation links 3D structure and genomic compartmentalization to CRD function and confirms our result that CRDs, in relation to TADs, are transcriptionally active intra-TAD domains. By losing CRD functionality through the loss of 3D structure it verifies that while some TSS are indeed induced similarity to wild-type transcription the other adjacent TSS are diminished or adversely effected in their activity. This key finding gives evidence of potential TSS communication perhaps due to the disruption of the macromolecule landscape that would otherwise be intact within the vicinity of these TSS. One limitation to this analysis is the departure from our RAW264.7 + LPS model system as well as our tight time interval time course highlighting the robust transcriptional response, in contrast HCT116 cells do not respond with the same magnitude. Our HCT116 + poly(I:C) time course being limited to 4 larger interval timepoints also diminished our temporal resolution.

It would be prudent, if developed, that a similar RAD21 degron or cohesin depletion experimental approach be deployed within our model system. Key insight as to which regulatory elements are diminished in the absence of normal 3D compartmentalization could identify genomic attributes that are optimal for maintaining transcription factor recruitment and normal gene expression levels (Rao et al., 2017). It is encouraging that, while departing from our model system, we observed CRD formation within a different species, cell type and stimulus. We also verify CRD characteristics and attributes discussed in chapter 3 within our HCT116 analysis.

One other key insight we highlight here is not only how natural genetic variation can be leveraged but also how potential CRD functions and mechanisms can be perturbed through mutagenesis. Here we report that when examining csRNA-seq profiled induction of C57BL/6 BMDM treated with LPS for 1hr vs SPRET BMDM of the same treatment we observe reduction not only at TSS with mutated NF-κB sites but non-mutated adjected TSS are also affected. While further examination would be prudent to verify this phenotype, it represents key evidence of regulatory elements communicating their induction to adjacent regulatory elements due to proximity. While proximity here is strictly limited and observed in linear genomic space this proximity communication could also extend to regulatory elements whose physical distance is reduced within the 3D genome space to due chromatin remodeling. However, to observe adjacent regulatory elements transcription induction greatly diminished simply due to proximity to a mutated NF-κB site in SPRET but non mutated and functional in C57BL/6 cells represents the elusive evidence into in-vivo regulatory element communication. To verify within our model system, we also observed the same dynamics when examining SPRET where mutations resulted in stronger consensus NF-κB motifs relative to C57BL/6.

From this analysis we report a crucial dynamic: if an NF-κB site is mutated and no other NF-κB site is within the region surrounding the TSS then the adjacent sites show diminished transcription induction within BMDM exposed to LPS. To add to this if we detect a wild-type NF-κB site adjacent to a candidate mutated NF-κB site within 50000bp the region in question doesn't show a similar loss of induction activity for adjacent TSS. This suggests the severity of a mutation could be defined by the presence of or the lack of other functional adjacent sites. With respect to CRDs and characterizing their functional components we report that the high saturation of NF-κB sites within their domain boundaries was correlated with induction of that region, the lower saturation identified a state of being downregulated. Taken together these findings contribute evidence into CRD function and regulatory element communication.

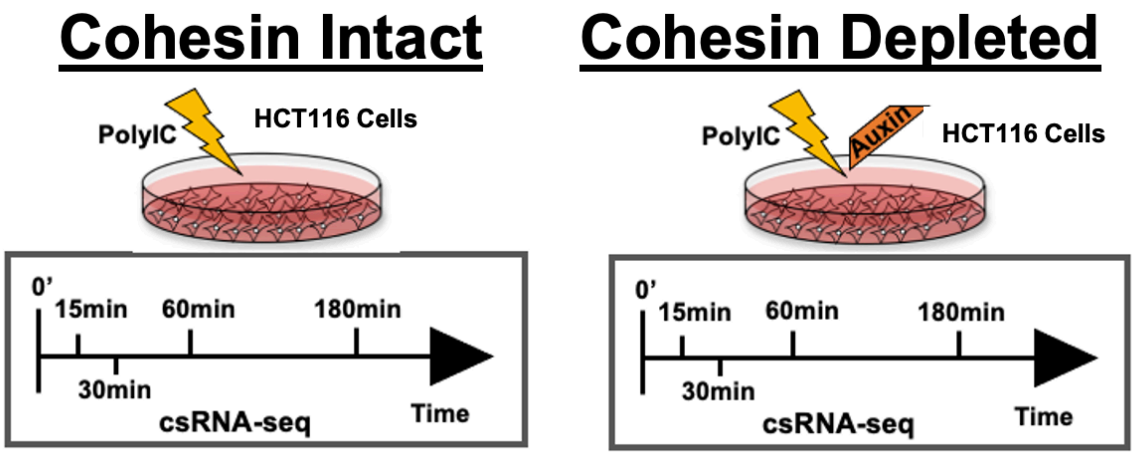

Figure 5.1 Overview of the HCT116 dataset, with treatment type and time points when samples were curated.

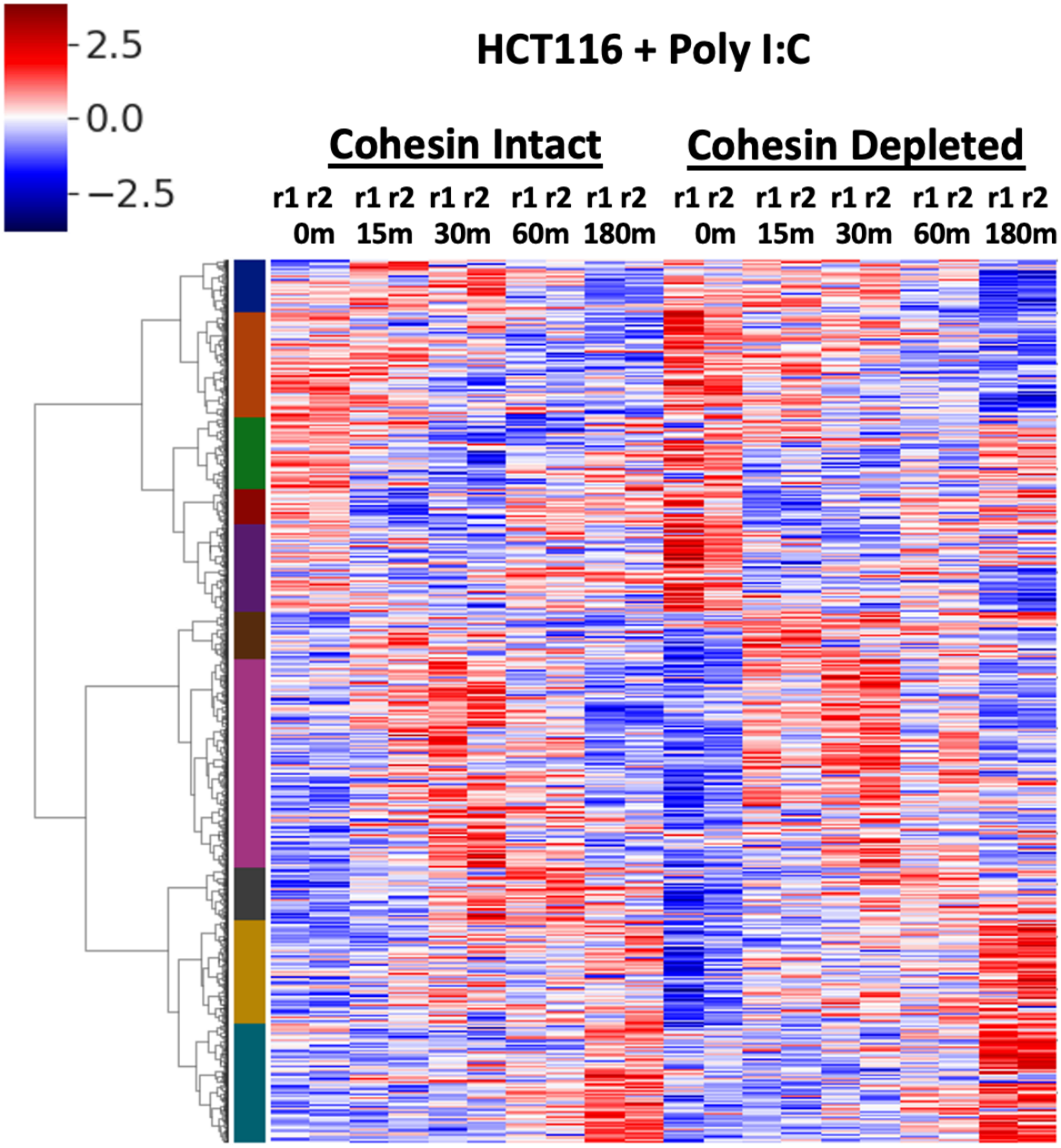

Figure 5.2 The HCT116 + Poly I:C time course heatmap representing hierarchically clustered transcription heatmap for cells with cohesin intact vs dysregulated.

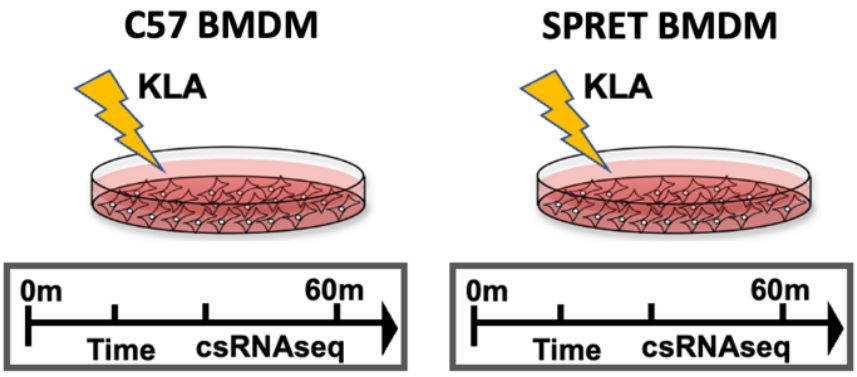

Figure 5.3 Overview of the two-time series datasets for C57BL/6 and SPRET mouse strains BMDM, with treatment type and time points when samples were extracted.

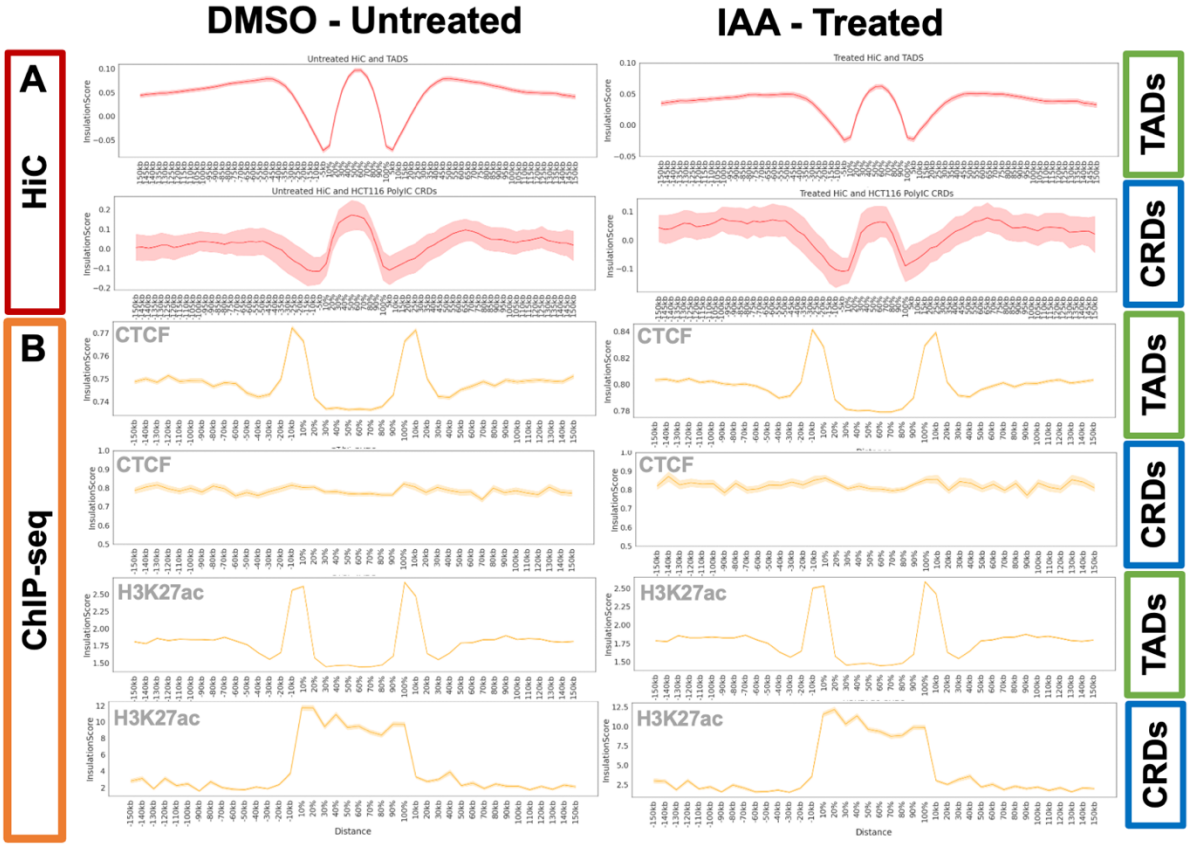

Figure 5.4 Hi-C and ChIP-seq results examining CRD and TAD characteristics within HCT116 cells after LPS exposure.

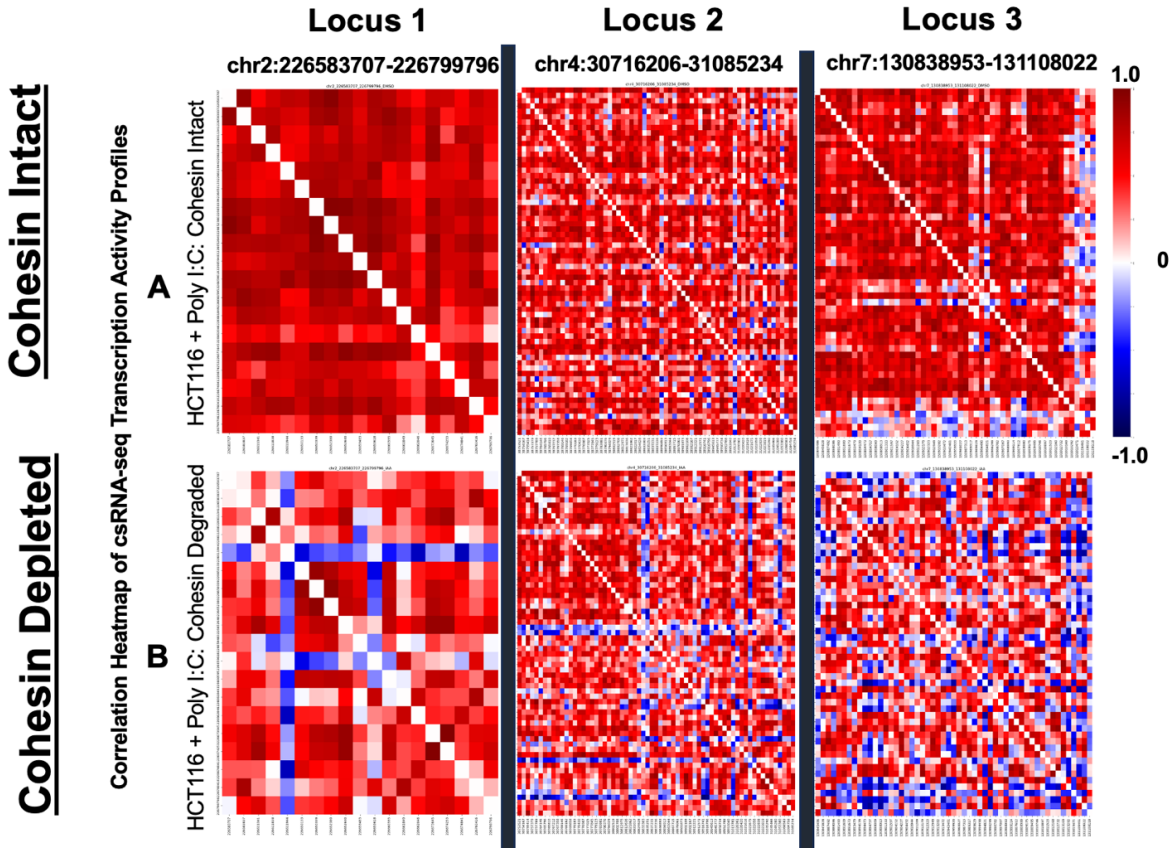

Figure 5.5 CRD TSS correlation heatmaps of specific genomic regions within HCT116 cells comparing correlation between (A) cells with cohesin intact vs (B) cells where cohesin is dysregulated.

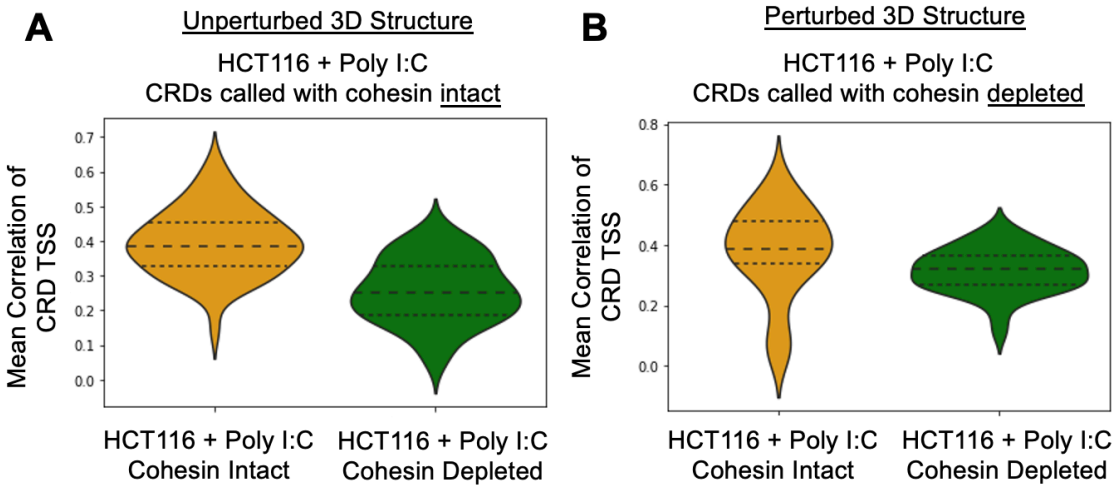

Figure 5.6 Mean CRD TSS correlation values within CRDs called within cells with cohesin intact and with cells where cohesin is dysregulated. TSS included are those shared between both datasets.

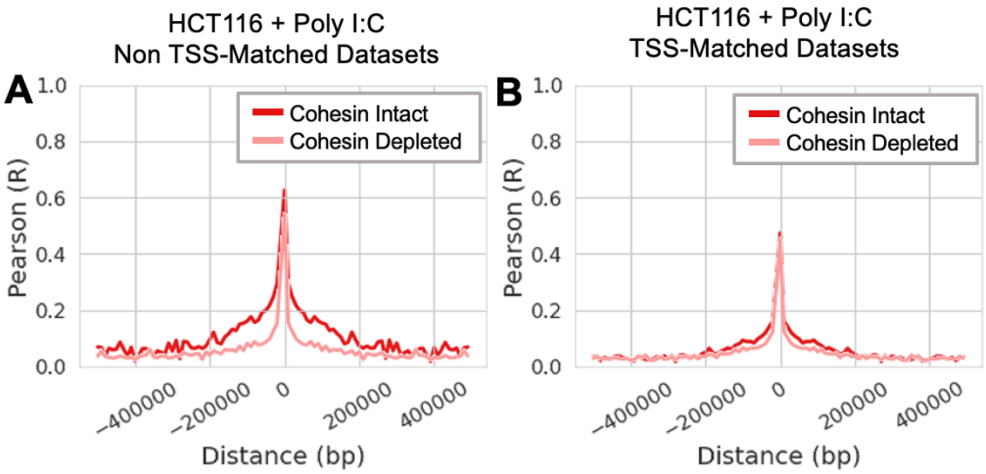

Figure 5.7 Line plots representing correlation between TSS and their adjacent TSS in with cohesin intact or with cohesin dysregulated. (A) TSS identified specifically from their respective dataset. (B) TSS that intersect between the two datasets.

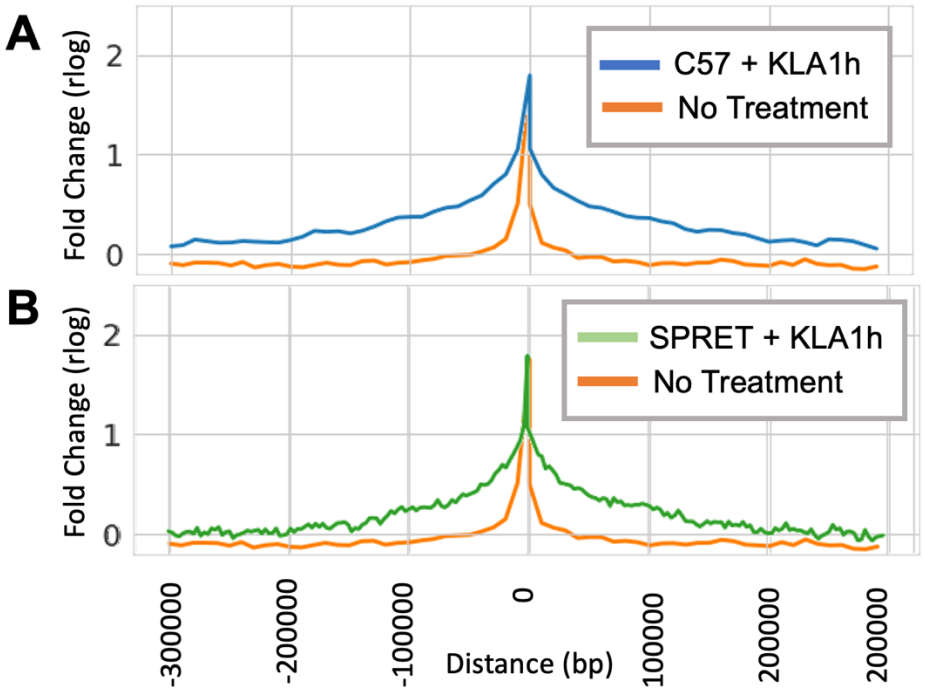

Figure 5.8 TSS Fold Changes at 1hr at C57BL/6 and SPRET TSS relative to a no treatment baseline curated between both C57BL/6 and SPRET TSS.

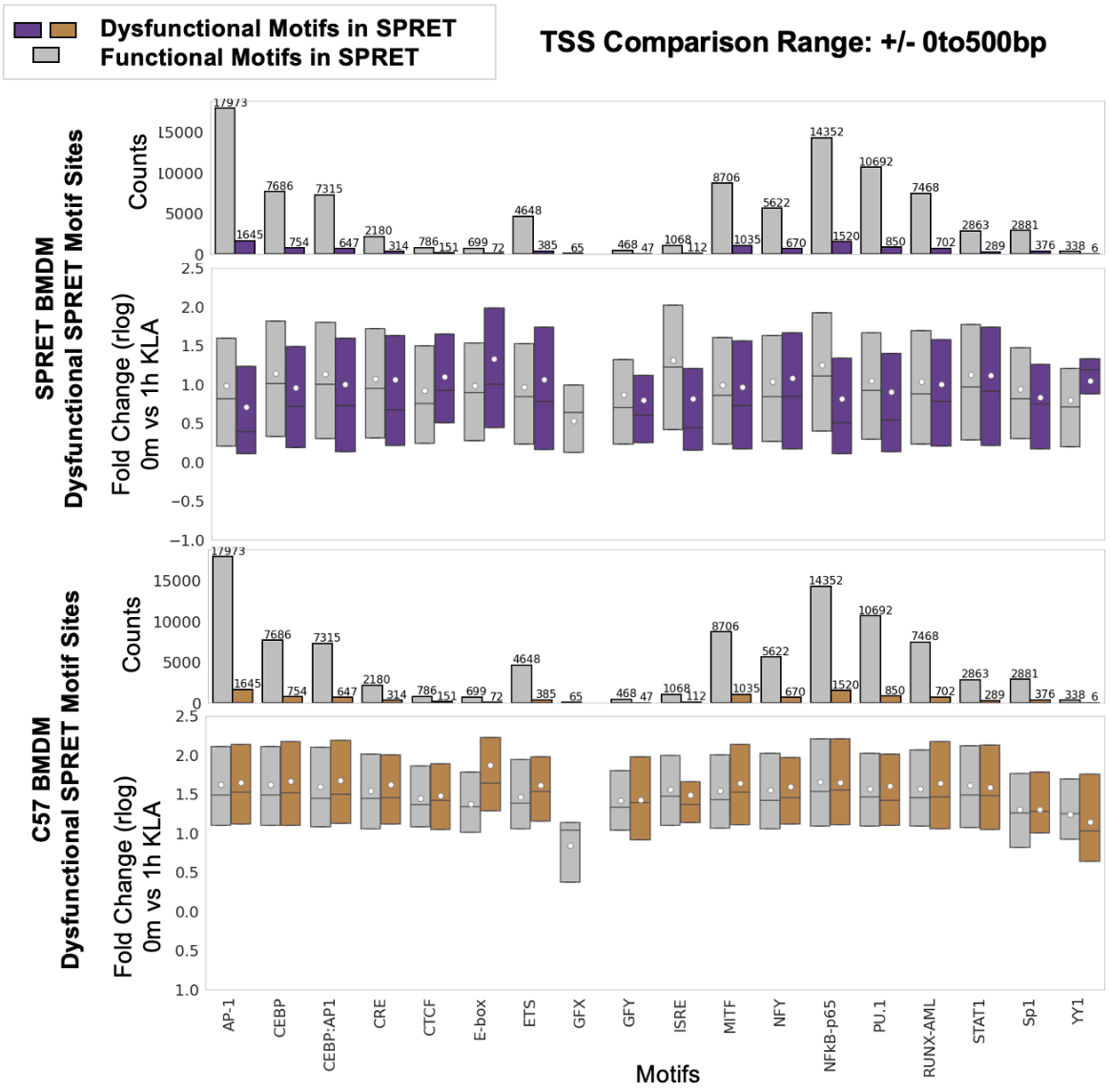

Figure 5.9 TSS Induction Levels of within C57BL/6 and SPRET mice comparing functional and dysfunctional motifs at SPRET sites, range limited to TSS +/- 0-500bp from candidate TSS.

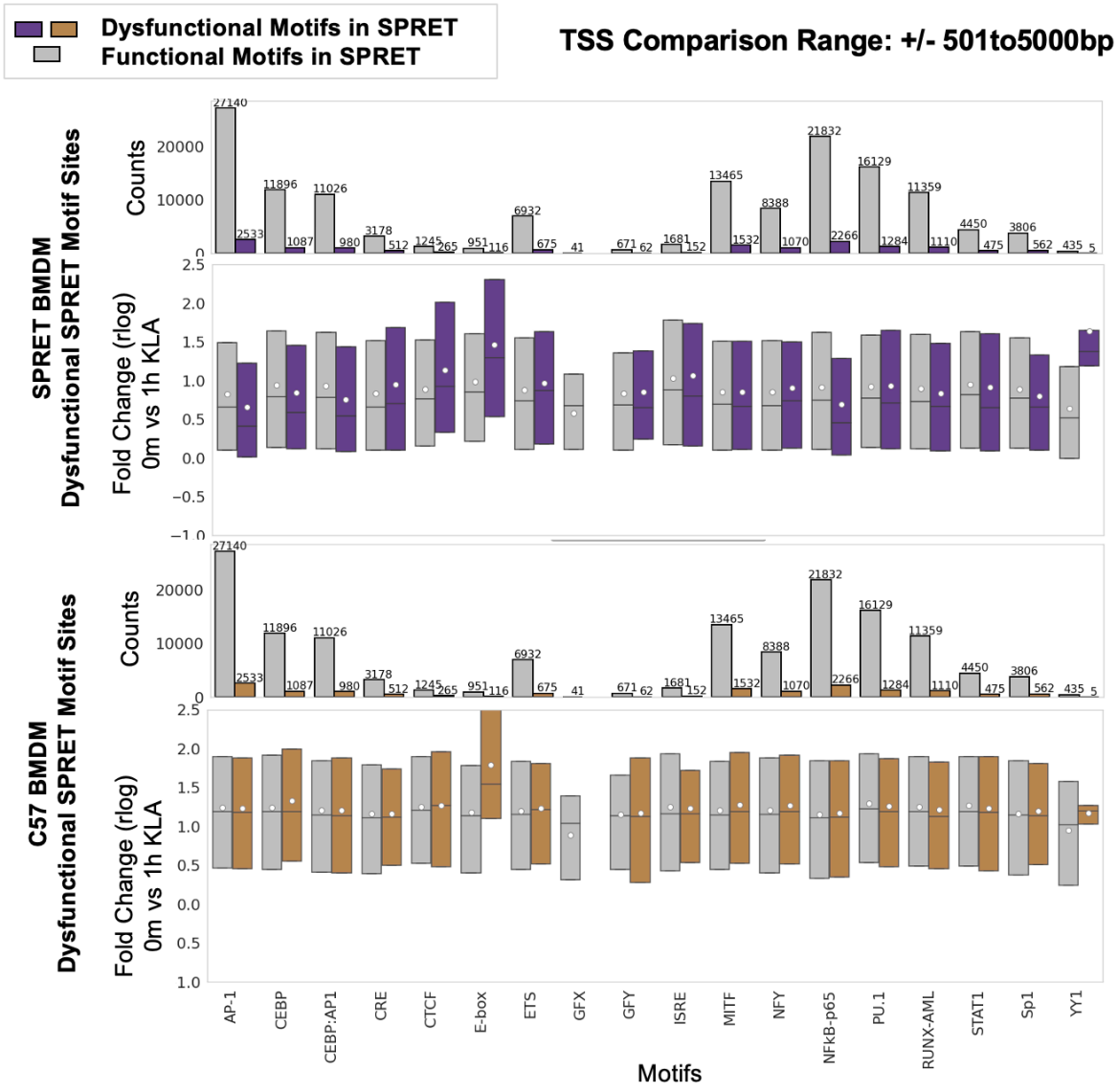

Figure 5.10 TSS Induction Levels of within C57BL/6 and SPRET mice comparing functional and dysfunctional motifs at SPRET sites, range limited to TSS +/- 501-5000bp from candidate TSS.

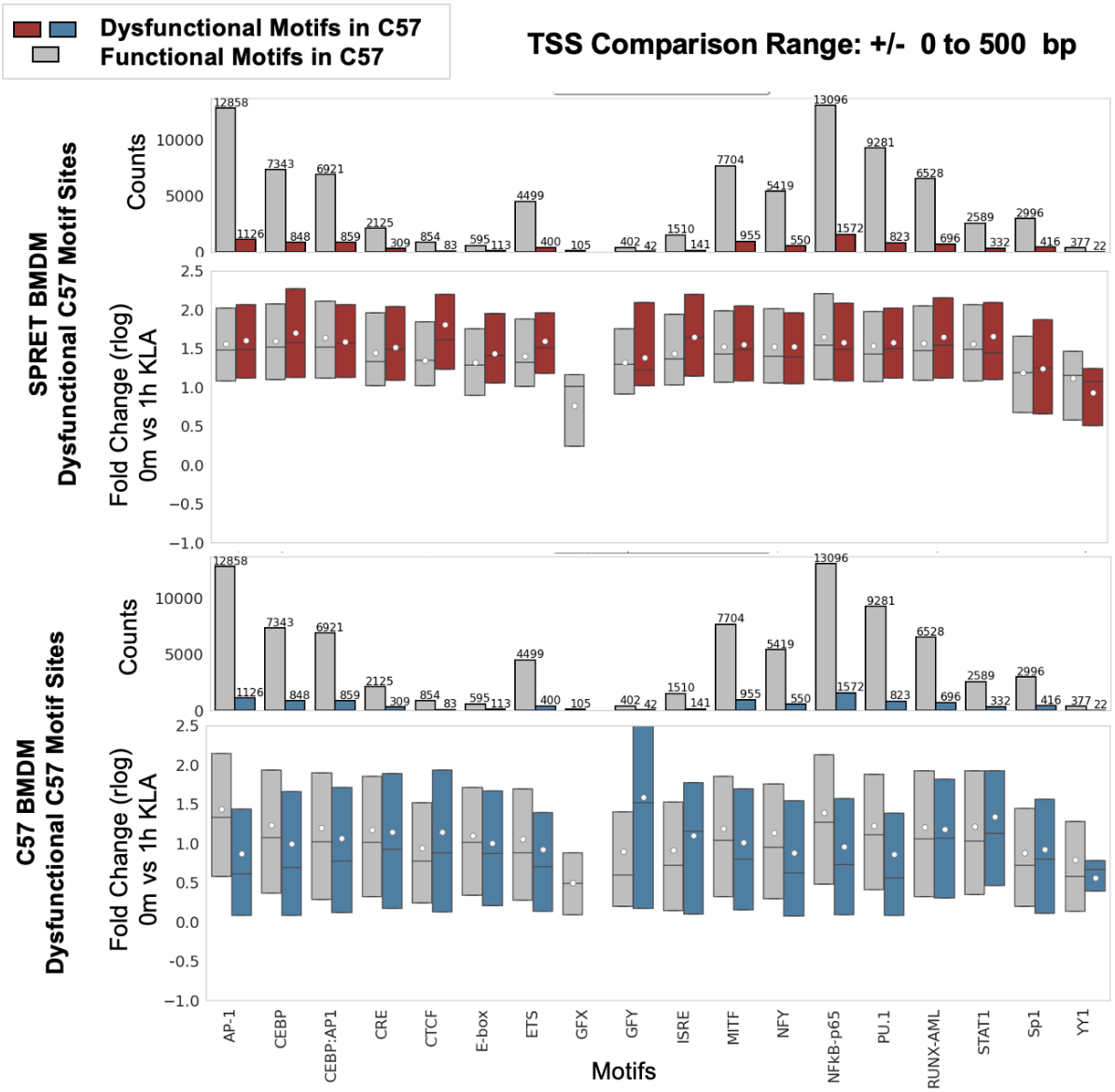

Figure 5.11 TSS Induction Levels of within C57BL/6 and SPRET mice comparing functional and dysfunctional motifs at C57BL/6 sites, range limited to TSS +/- 0-500bp from candidate TSS.

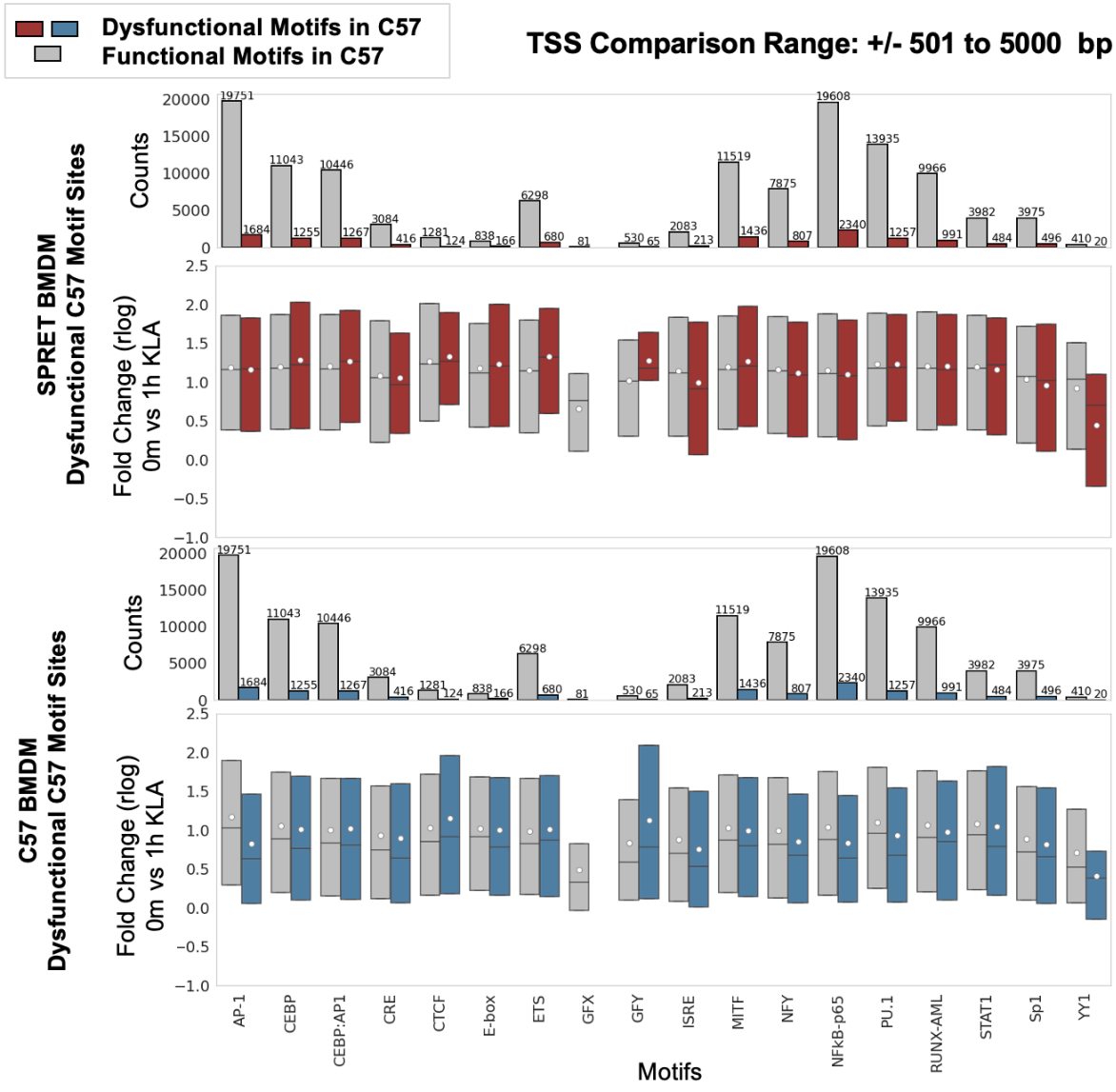

Figure 5.12 TSS Induction Levels of within C57BL/6 and SPRET mice comparing functional and dysfunctional motifs at C57BL/6 sites, range limited to TSS +/- 501-5000bp from candidate TSS.

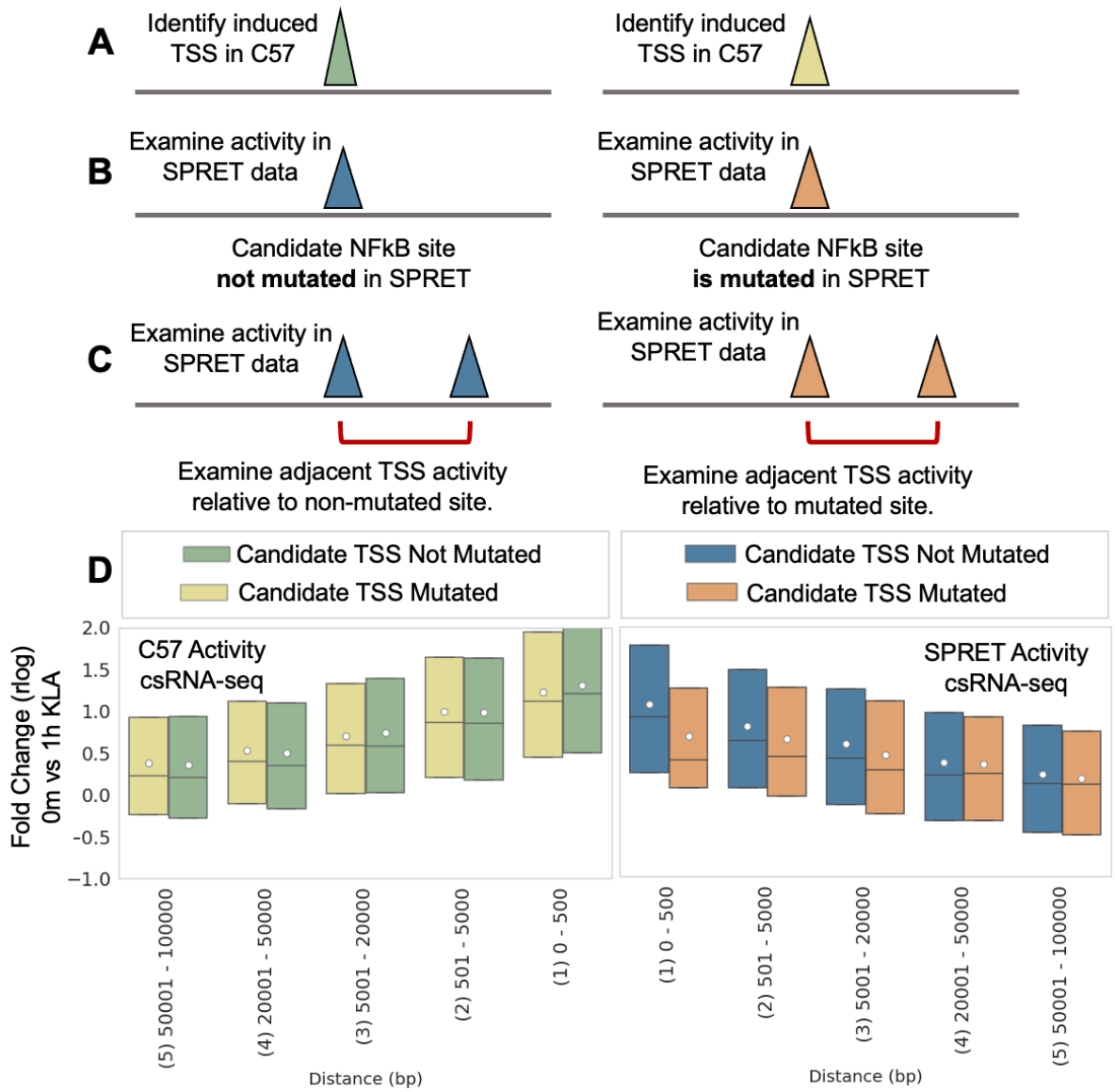

Figure 5.13 Schematic and results examining mutated or non-mutated NF-κB SPRET TSS. (A) The selection criteria. (B) The data and color representation of the SPRET datasets. (C) TSS data collection criteria for candidate and adjacent TSS. (D) Binned TSS activity for C57BL/6 and SPRET TSS at mutated or non-mutated NF-κB SPRET TSS.

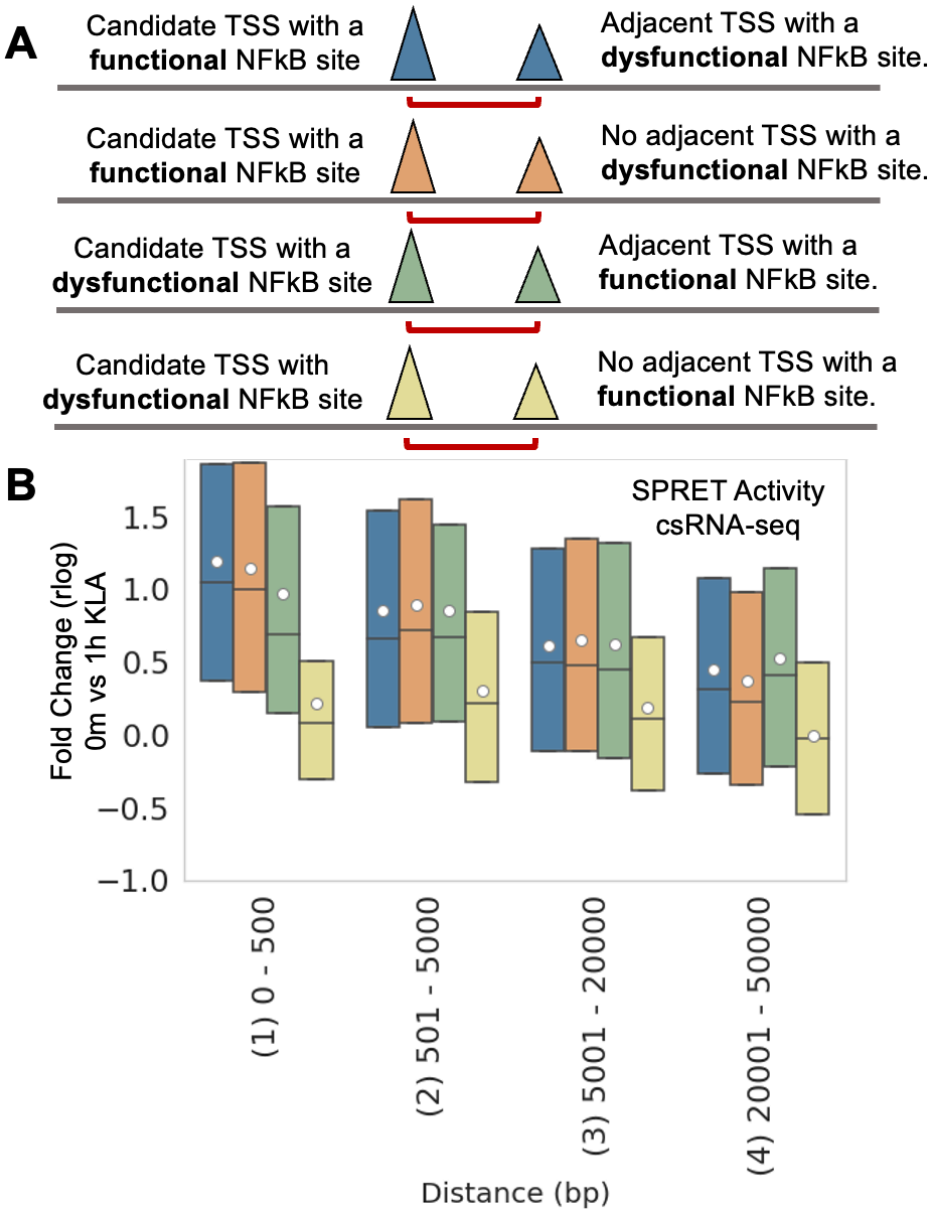

Figure 5.14 (A) Schematic and (B) results examining NF-κB mutations at SPRET TSS with additional annotations as to if adjacent TSS are also mutated or non-mutated NF-κB TSS.

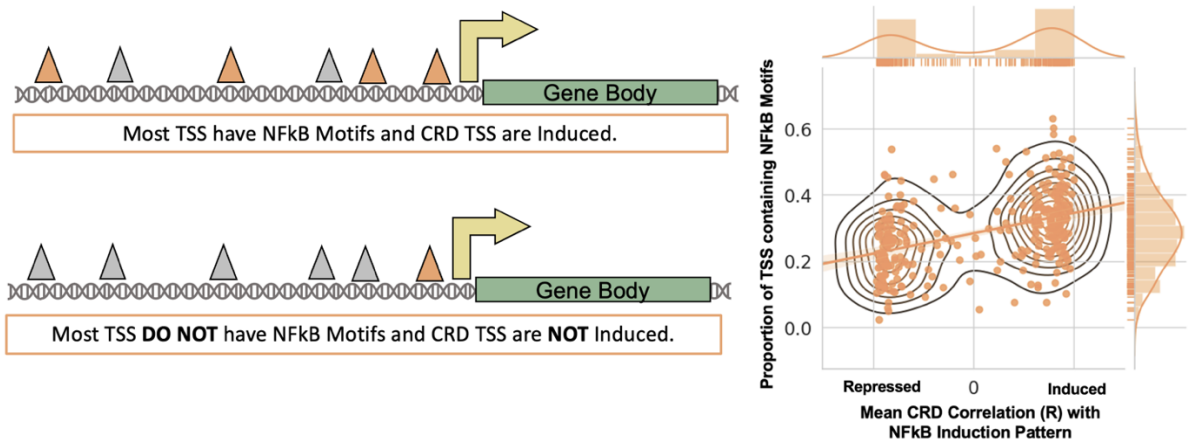

Figure 5.15 Mean CRD correlation with NF-κB induction activity relative to the proportion of CRD TSS containing NF-κB transcription factor binding motifs.

Acknowledgements

Chapters 1,2,4,5,6 represent research intended to be published together pending further insight provided by a Dual-MPRA approach described in Chapter 6. Hillary, R., Guzman, C., Heinz, S., Benner, C. The dissertation author will be the primary investigator and author of this paper.

Chapter 6 Conclusions

This study demonstrated the versatility of nascent transcription profiling techniques and their ability to examine transcription dynamics at a high temporal and spatial resolution (Duttke et al., 2019). This data-driven study sought to address an elusive question: if and how regulatory elements communicate with one another to drive changes in the local regulatory landscape.

As we followed a strictly data driven research strategy many aspects of transcription dynamics were examined and characterized in this study. Within our model system we demonstrate that csRNA-seq, when applied within a time series approach, identifies key windows of regulatory moments. This regulatory hierarchy within LPS treated macrophages shows a rapid induction and

usage of NF-κB containing regulatory elements at both enhancers and promoters, immediately followed by a strong induction of AP-1 usage that appears intwined with NF-κB dynamics. This strong initial transcriptional burst is followed by regulatory elements housing IRF transcription factor binding motifs. This continued characterization of these precise windows of regulatory dynamics allowed us to directly identify and focus further analyses into the first, initial burst of transcription within our model system.

Focusing on the strong initial burst of transcription within LPS-treated macrophages allowed key insight into a critical distinction between enhancer and promoter induction dynamics. While previously reported that enhancer transcription induction precedes promoters, we clarified that specifically within our robust stimulus driven approach the induction of enhancers and promoters is nearly simultaneous. This insight is key in moving forward to better deconvolute true precursors driving transcription factor recruitment and transcription initiation. We found that, while the mean induction times between enhancers and promotors was nearly simultaneous even when examined within a tight time interval (2 min) time course, some loci can have promoters precede enhancers and vice versa. This dynamic suggests that transcription recruitment is more focused on regulatory element motif composition and concludes that enhancer/promoter distinction alone isn't a clear means to determine which regulatory element precedes another in induction timing.

By characterizing that induction between localized regulatory elements is near simultaneous we also report that these localized sites of intense regulatory dynamics form domains of correlated transcription activity. We presented an algorithm to identify and extract these domains to elucidate their prevalence genome wide and assess their function as a genomic entity. We detected 153 of these cis-regulatory domains that are comprised of TSS among the first to

respond to our LPS stimulus within macrophages and are primarily enriched for NF-κB motifs relative to down regulated CRD TSS. We established that these domains do have characteristics and genomic attributes to suggest they are their own genuine genomic entities and not either super enhancers or just TADs. In relation to TADs, CRDs are transcriptionally active intra-TAD domains as most CRDs appear within sub-TAD compartments when examining 3D structure characteristics, even after examining a resolution down to 10k within a deeply sequenced Hi-C dataset.

We validated our CRD characteristics in another species, with other stimuli and cell types by leveraging a timeseries dataset using csRNA-seq examining transcription dynamics within HCT116 cells exposed to poly(I:C). Here, we took advantage of a HCT116 with a genetically engineered auxin-inducible RAD21 degron system to examine CRD formation in the absence of cohesin. (Rao et al., 2017) The reduction of cohesin and significant loss of 3D structure, by means of perturbing genomic looping and TAD boundary dysregulation, resulted in highly degraded CRDs containing individual TSS that did not share the characteristic correlated transcription activity profiles as seen in unperturbed cells. This established that CRD function and correlated TSS activity within their domains are associated with intact 3D structure, such as TADs and genomic looping. This conclusion also suggests that the loss of 3D structure could also impede potential regulatory element communication. Further work is necessary to solidify these results.

We reported that the loss of NF-κB binding affinity of SPRET TSS due to natural genetic variation relative to C57BL/6 (Keane et al., 2011) revealed not only the loss of transcription but also transcription activity of adjacent non- NF-κB motif-containing regulatory elements. This result is encouraging evidence for potential regulatory element/regulatory element communication, an elusive mechanism that will be crucial to understand if regulatory elements are

independent or dependent on other, separate TSS for their induction and transcription activity. We also conclude that this mutation of NF-κB sites and their effect on adjacent sites is dependent on whether or not another functional NF-κB site is located nearby. If no other functional NF-κB site is near, this loss of transcription can extend up to 50kb away. This dynamic adds additional insight into assessing mutation severity but also into CRD functionality in general as we suspected that CRDs are group of sites containing key transcription factor binding motifs recruiting for the entire domain. This insight could also help with the interpretation of studies mutating individual sites in-vivo and their effect on overall transcription and associated gene expression levels, as single mutations could be nullified but functional adjacent TSS. We report the association of high saturation of NF-κB-containing TSS with induction potential of the CRD adding to what makes a CRD a CRD: a collection of TSS containing associated transcription factor binding motifs, recruiting and communicating transcription activity to adjacent sites, including promoter sites.

Future studies examining CRDs would be prudent to solidify these conclusions. It will be important examine the function of CRD TSS outside of their genomic context. One future study would be needed to identify if indeed some regulatory elements were only induced in their genomic context as not all induced CRD TSS contained TLR4 associated motifs yet were induced. We hypothesize that these sites were only induced due to direct proximity to another TSS that contained such motifs and were the primary recruiters of transcription factors, RNA polymerase II and other proteins. To establish this dynamic would clarify the true recruiters of transcription within CRDs and if it is indeed all or just a subset of induced TSS. Examining transcription within living cells is indeed a daunting process given the multitude of regulatory signals identified examining nascent RNA transcripts. To further refine analyses to directly identify which TSS are either targeted by or directly recruiting transcription factors would help to identify and reduce

potential noise. For this study we could leverage massively parallel reporter assays (MPRA), specifically a Dual-MPRA (Carlos Guzmán, Ph.D. thesis 2023) which permits the inclusion and measure of two individual TSS activity on a single plasmid. This approach would answer two questions: does a TSS require its context to be induced and do adjacent TSS influence another TSS activity. The TSS for this were selected from all CRD TSS and mutagenesis was also performed on a subset of these TSS to further solidify developing motif grammar rules. The future completion of this study will be vital to provide further evidence of regulatory element communication.

Acknowledgements

Chapters 1,2,4,5,6 represent research intended to be published together pending further insight provided by a Dual-MPRA approach described in Chapter 6. Hillary, R., Guzman, C., Heinz, S., Benner, C. The dissertation author will be the primary investigator and author of this paper.

REFERENCES

Alexander, J. M., Guan, J., Li, B., Maliskova, L., Song, M., Shen, Y., … Weiner, O. D. (2019). Live-cell imaging reveals enhancer-dependent Sox2 transcription in the absence of enhancer proximity. ELife, 8, e41769. https://doi.org/10.7554/eLife.41769

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., … Bagger, F. O. (2014). An atlas of active enhancers across human cell types and tissues. Nature, 507(7493), 455–461. https://doi.org/10.1038/nature12787

Arner, E., Daub, C. O., Vitting-Seerup, K., Andersson, R., Lilje, B., Drabløs, F., … Davis, M. (2015). Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science, 347(6225), 1010–1014. https://doi.org/10.1126/science.1259418

Blackwood, E. M. (1998). Going the Distance: A Current View of Enhancer Action. Science, 281(5373), 60–63. https://doi.org/10.1126/science.281.5373.60

Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Current Protocols in Molecular Biology, 21.29.1–21.29.9. https://doi.org/10.1002/0471142727.mb2129s109

Bulger, M., & Groudine, M. (1999). Looping versus linking: toward a model for long-distance gene activation. Genes & Development, 13(19), 2465–2477. https://doi.org/10.1101/gad.13.19.2465

Cheng, C. S., Behar, M., Suryawanshi, G. W., Feldman, K. E., Spreafico, R., & Hoffmann, A. (2017). Iterative Modeling Reveals Evidence of Sequential Transcriptional Control Mechanisms. 4(3), 330-343.e5. https://doi.org/10.1016/j.cels.2017.01.012

Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., & Lis, J. T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature Genetics, 46(12), 1311–1320. https://doi.org/10.1038/ng.3142

Ding, G., Fischer, P., Boltz, R. C., Schmidt, J., Colaianne, J. J., Gough, A., … Miller, D. C. (1998). Characterization and Quantitation of NF-κB Nuclear Translocation Induced by Interleukin-1 and Tumor Necrosis Factor-α. *Journal of Biological Chemistry*, *273*(44), 28897–28905. https://doi.org/10.1074/jbc.273.44.28897

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2012). STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Duttke, S. H., Chang, M. W., Heinz, S., & Benner, C. (2019). Identification and dynamic quantification of regulatory elements using total RNA. Genome Research, 29(11), 1836–1846. https://doi.org/10.1101/gr.253492.119

Fang, W., Bi, D., Zheng, R., Cai, N., Xu, H., Zhou, R., … Xu, X. (2017). Identification and activation of TLR4-mediated signalling pathways by alginate-derived guluronate oligosaccharide in RAW264.7 macrophages. Scientific Reports, 7(1). https://doi.org/10.1038/s41598-017-01868-0

Ferreiro, D. U., & Komives, E. A. (2010). Molecular Mechanisms of System Control of NF-κB Signaling by IκBα. Biochemistry, 49(8), 1560–1567. https://doi.org/10.1021/bi901948j

Field, A., & Adelman, K. (2020). Evaluating Enhancer Function and Transcription. Annual Review of Biochemistry, 89(1), 213–234. https://doi.org/10.1146/annurev-biochem-011420-095916

Fujioka, S., Niu, J., Schmidt, C., Sclabas, G. M., Peng, B., Uwagawa, T., … Chiao, P. J. (2004). NF-κB and AP-1 Connection: Mechanism of NF-κB-Dependent Regulation of AP-1 Activity. Molecular and Cellular Biology, 24(17), 7806–7819. https://doi.org/10.1128/mcb.24.17.7806-7819.2004

Gower, J. C. (1985). Properties of Euclidean and non-Euclidean distance matrices. Linear Algebra and Its Applications, 67, 81–97. https://doi.org/10.1016/0024-3795(85)90187-9

Guzmán, C., Duttke, S. H., Zhu, Y., De, C., Downes, N., Benner, C., & Heinz, S. (2023). Combining TSS-MPRA and sensitive TSS profile dissimilarity scoring to study the sequence determinants of transcription initiation. Nucleic Acids Research, 51(15), e80–e80. https://doi.org/10.1093/nar/gkad562

Hah, N., Murakami, S., Nagari, A., Danko, C. G., & Kraus, W. L. (2013). Enhancer transcripts mark active estrogen receptor binding sites. *Genome Research*, *23*(8), 1210–1223. https://doi.org/10.1101/gr.152306.112

Hargreaves, D. C., Horng, T., & Medzhitov, R. (2009). Control of Inducible Gene Expression by Signal-Dependent Transcriptional Elongation. Cell, 138(1), 129–145. https://doi.org/10.1016/j.cell.2009.05.047

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., … Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell, 38(4), 576–589. https://doi.org/10.1016/j.molcel.2010.05.004

Jindal, G. A., & Farley, E. K. (2021). Enhancer grammar in development, evolution, and disease: dependencies and interplay. Developmental Cell, 56(5), 575–587. https://doi.org/10.1016/j.devcel.2021.02.016

Karnuta, J. M., & Scacheri, P. C. (2018). Enhancers: bridging the gap between gene control and human disease. Human Molecular Genetics, 27(R2), R219–R227. https://doi.org/10.1093/hmg/ddy167

Kawasaki, T., & Kawai, T. (2014). Toll-Like Receptor Signaling Pathways. Frontiers in Immunology, 5(461). https://doi.org/10.3389/fimmu.2014.00461

Keane, T. M., Goodstadt, L., Danecek, P., White, M. A., Wong, K., Yalcin, B., … Oliver, P. L. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature, 477(7364), 289–294. https://doi.org/10.1038/nature10413

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., … Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature, 465(7295), 182–187. https://doi.org/10.1038/nature09033

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., … Carninci, P. (2006). CAGE: cap analysis of gene expression. Nature Methods, 3(3), 211–222. https://doi.org/10.1038/nmeth0306-211

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

Levine, M. (2010). Transcriptional enhancers in animal development and evolution. Current Biology: CB, 20(17), R754-763. https://doi.org/10.1016/j.cub.2010.06.070

Li, G., Hao, W., & Hu, W. (2020). Transcription factor PU.1 and immune cell differentiation (Review). *International Journal of Molecular Medicine*, *46*(6), 1943–1950. https://doi.org/10.3892/ijmm.2020.4763

Li, J., Hsu, A., Hua, Y., Wang, G., Cheng, L., Ochiai, H., … Pertsinidis, A. (2020). Single-gene imaging links genome topology, promoter–enhancer communication and transcription control. Nature Structural & Molecular Biology, 27(11), 1032–1040. https://doi.org/10.1038/s41594-020-0493-6

Link, V. M., Duttke, S. H., Chun, H. B., Holtman, I. R., Westin, E., Hoeksema, M. A., … Ren, B. (2018). Analysis of Genetically Diverse Macrophages Reveals Local and Domain-wide Mechanisms that Control Transcription Factor Binding and Function. Cell, 173(7), 1796-1809.e17. https://doi.org/10.1016/j.cell.2018.04.018

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12). https://doi.org/10.1186/s13059-014-0550-8

Lovén, J., Hoke, Heather A., Lin, Charles Y., Lau, A., Orlando, David A., Vakoc, Christopher R., … Young, Richard A. (2013). Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers. Cell, 153(2), 320–334. https://doi.org/10.1016/j.cell.2013.03.036

McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., … Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nature Biotechnology, 28(5), 495–501. https://doi.org/10.1038/nbt.1630

Michel, M., Demel, C., Zacher, B., Schwalb, B., Krebs, S., Blum, H., … Cramer, P. (2017). *TT-seq captures enhancer landscapes immediately after T-cell stimulation*. *13*(3), 920–920. https://doi.org/10.15252/msb.20167507

Mikhaylichenko, O., Bondarenko, V., Harnett, D., Schor, I. E., Males, M., Viales, R. R., & Furlong, E. E. M. (2018). The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. Genes & Development, 32(1), 42–57. https://doi.org/10.1101/gad.308619.117

Nanni, L., Ceri, S., & Logie, C. (2020). Spatial patterns of CTCF sites define the anatomy of TADs and their boundaries. Genome Biology, 21(1). https://doi.org/10.1186/s13059-020-02108-x

Nezar Abdennur, Abraham, S., Fudenberg, G., Flyamer, I. M., Galitsyna, A. A., Goloborodko, A., … Venev, S. V. (2022). Cooltools: enabling high-resolution Hi-C analysis in Python. BioRxiv (Cold Spring Harbor Laboratory). https://doi.org/10.1101/2022.10.31.514564

Olivier Delaneau, Zazhytska, M., Borel, C., Giannuzzi, G., Rey, G., Howald, C., … Antonarakis, S. E. (2019). Chromatin three-dimensional interactions mediate genetic effects on gene expression. Science, 364(6439). https://doi.org/10.1126/science.aat8266

Pott, S., & Lieb, J. D. (2015). What are super-enhancers? *Nature Genetics*, *47*(1), 8–12. https://doi.org/10.1038/ng.3167

Rao, S. S. P., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K.-R., … Casellas, R. (2017). Cohesin Loss Eliminates All Loop Domains. Cell, 171(2), 305-320.e24. https://doi.org/10.1016/j.cell.2017.09.026

Rickels, R., & Shilatifard, A. (2018). Enhancer Logic and Mechanics in Development and Disease. Trends in Cell Biology, 28(8), 608–630. https://doi.org/10.1016/j.tcb.2018.04.003

Ryan, G. E., & Farley, E. K. (2019). Functional genomic approaches to elucidate the role of enhancers during development. WIREs Systems Biology and Medicine, 12(2). https://doi.org/10.1002/wsbm.1467

Sabari, B. R., Dall'Agnese, A., Boija, A., Klein, I. A., Coffey, E. L., Shrinivas, K., … Roeder, R. G. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. Science, 361(6400), eaar3958. https://doi.org/10.1126/science.aar3958

Sergei Nechaev, Fargo, D. C., Lima, G., Liu, L., Gao, Y., & Adelman, K. (2010). Global Analysis of Short RNAs Reveals Widespread Promoter-Proximal Stalling and Arrest of Pol II in Drosophila. Science, 327(5963), 335–338. https://doi.org/10.1126/science.1181421

Thakore, P. I., D'Ippolito, A. M., Song, L., Safi, A., Shivakumar, N. K., Kabadi, A. M., … Gersbach, C. A. (2015). Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. Nature Methods, 12(12), 1143–1149. https://doi.org/10.1038/nmeth.3630

Tong, A.-J., Liu, X., Thomas, B. J., Lissner, M. M., Baker, M. R., Senagolage, M. D., … Smale, S. T. (2016). A Stringent Systems Approach Uncovers Gene-Specific Mechanisms Regulating Inflammation. Cell, 165(1), 165–179. https://doi.org/10.1016/j.cell.2016.01.020

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association, 58(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., … Young, R. A. (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell, 153(2), 307–319. https://doi.org/10.1016/j.cell.2013.03.035

Wolff, J., Rabbani, L., Gilsbach, R., Richard, G., Manke, T., Backofen, R., & Grüning, B. A. (2020). Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. Nucleic Acids Research, 48(W1), W177–W184. https://doi.org/10.1093/nar/gkaa220

Zhang, Z., Lee, J.-H., Ruan, H., Ye, Y., Krakowiak, J., Hu, Q., … Han, L. (2019). Transcriptional landscape and clinical utility of enhancer RNAs for eRNA-targeted therapy in cancer. Nature Communications, 10(1). https://doi.org/10.1038/s41467-019-12543-5

Zhao, B., Barrera, L. A., Ersing, I., Willox, B., Stefanie Julia Schmidt, Greenfeld, H., … Gewurz, B. E. (2014). The NF-κB Genomic Landscape in Lymphoblastoid B Cells. Cell Reports, 8(5), 1595–1606. https://doi.org/10.1016/j.celrep.2014.07.037