

# The Dissection of VEGFA Stimulus-Responsive Regulatory and Transcriptional Changes in Angiogenesis

by

Daniel S. Day

B.A., Augustana College (2009)

Submitted to the Harvard-MIT Program in Health Sciences and Technology

in partial fulfillment of the requirements for the degree of

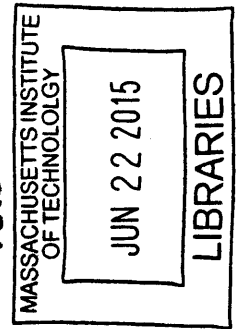
Doctor of Philosophy in Medical Engineering Medical Physics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

©Massachusetts Institute of Technology, 2015. All rights reserved.



**Signature redacted**

Author .....  
Harvard-MIT Program in Health Sciences and Technology

May 18, 2015

**Signature redacted**

Certified by ..  
.....  
Peter J. Park

Professor of Pediatrics  
Thesis Supervisor

**Signature redacted**

Accepted by ..  
.....  
Emery Brown

MD, PhD/Director, Harvard-MIT Program in Health Sciences and Technology/Professor of Computational Neuroscience and Health Sciences and Technology

# **The Dissection of VEGFA Stimulus-Responsive Regulatory and Transcriptional Changes in Angiogenesis**

by  
Daniel S. Day

Submitted to the Harvard-MIT Program in Health Sciences and Technology  
on May 18, 2015, in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy in Medical Engineering Medical Physics

## **Abstract**

Several studies over the past decade have transformed our understanding of the regulatory elements and mechanisms utilized by a human cell to drive cell type identity. In particular, epigenomic studies have revealed recurrent epigenetic signatures at enhancers and other regulatory regions, as well as their role in cellular lineage specification. However, these studies generally focused on steady-state cellular states where much of the lifespan of adult cells involves responding to extracellular cues. To better understand the gene expression changes that occur in response to stimuli, I studied a time-course stimulation of human umbilical vascular endothelial cells (HUVECs) with vascular endothelial growth factor A (VEGFA) as a model system. Using data collected from multiple genome-wide assays I modeled the dynamic changes in epigenetic, transcriptional, and transcription factor binding profiles in regulation of angiogenesis, the formation of new blood vessels. First, I identify regulatory elements involved in VEGFA-response through focal, temporal changes in chromatin structure and that p300 activity is mechanistically required for this response. Secondly, I analyze changes in combinatorial binding of transcription factors linked with VEGFA-responsive enhancers. These studies highlight general strategies to study stimulus-responsive regulatory systems, and reveal new insights into angiogenesis, human disease and therapeutic targets. Finally, I show that VEGFA-responsive genes are regulated by promoter-proximal RNA Polymerase II pausing and extend it to comprehensive analysis of gene expression and chromatin regulation by promoter-proximal pausing across cell types.

Thesis Supervisor: Peter J. Park  
Title: Professor of Pediatrics

## **Acknowledgments**

To my parents, sisters, brother, grandparents, aunts, uncles and cousins, I am forever grateful for all the love and support given to me over the years. I would not be here without all of you in my life.

To my friends, thank you for helping me celebrate the good times and get through the tough ones. I am so fortunate to have you all in order to keep be grounded and focused on what is important in life.

To my colleagues, thank you for your guidance, support and answers to my questions over the many years. Through all of your help experimentally, analytically, and even just sharing ideas has allowed me to grow into a better scientist.

# Contents

<b>List of Figures</b>	<b>7</b>
<b>List of Tables</b>	<b>11</b>
<b>1 Introduction</b>	<b>12</b>
1.1 The Composition of the Human Genome and the Types of Genes . . . . .	12
1.2 The Organization of Chromatin and the Nucleus . . . . .	13
1.3 Transcription Factors Bind Regulatory Elements and Regulate Transcription	22
1.4 Enhancers Regulate Gene Expression by Looping Towards Target Promoters	23
1.5 Chromatin Remodeling During Development Compared with Chromatin Remodeling During Signal Response . . . . .	24
1.6 The regulation of Pol II and transcription by regulating both initiation and the entry into elongation . . . . .	25
1.7 The Physiological Process of Creating New Blood Vessels: Angiogenesis .	26
1.8 Understanding the Chromatin Dynamics and Transcriptional Regulation of Angiogenesis . . . . .	27
<b>2 A dynamic H3K27ac signature identifies VEGFA stimulated endothelial en- hancers and requires p300 activity</b>	<b>29</b>
2.1 Abstract . . . . .	29
2.2 Introduction . . . . .	30
2.3 Results . . . . .	30
2.3.1 VEGFA Stimulation Induces Local H3K27ac Changes . . . . .	30
2.3.2 A dynamic VEGFA-regulated H3K27ac signature is tightly linked to p300 chromatin occupancy . . . . .	31
2.3.3 Temporal clustering of H3K27ac variation defined groups of chro- matin regions with distinct function annotations and enriched tran- scription factor motifs . . . . .	35
2.3.4 The dynamic H3K27ac signature defines VEGFA-responsive tran- scriptional regulatory elements . . . . .	41
2.3.5 Dynamic H3K27ac sites and p300 participate in VEGFA-stimulated chromatin looping . . . . .	43
2.4 Discussion . . . . .	44
2.5 Methods . . . . .	47
2.5.1 Experimental Work and Growth Conditions . . . . .	47



2.5.2	ChIP-seq Analysis, Peak Calling for p300 and DNase I Hypersensitivity Sites . . . . .	47
2.5.3	Calculating the H3K27ac variance score . . . . .	48
2.5.4	RNA-seq Analysis . . . . .	48
2.5.5	Transcription factor motif analysis . . . . .	48
2.5.6	Browser Views . . . . .	49
2.5.7	Data access . . . . .	49
<b>3</b>	<b>Analyzing the Transcriptional Regulation of Endothelial Cells and VEGFA Stimulus-Response</b>	<b>50</b>
3.1	Abstract . . . . .	50
3.2	Introduction . . . . .	50
3.3	Results . . . . .	51
3.3.1	Experimental Design . . . . .	51
3.3.2	Analysis of Temporal Pattern of VEGFA-responsive Protein Coding and Non-coding RNA Gene Expression Change . . . . .	51
3.3.3	Low Temporal Variability of Typical Chromatin Marks of Active TSSes at VEGFA Responsive Genes . . . . .	57
3.3.4	The Broad Chromatin Landscape is Stable under VEGFA Starvation and Stimulation . . . . .	62
3.3.5	The Enhancer Landscape during VEGFA Response . . . . .	63
3.3.6	Comparison of Active Regulatory Regions during VEGFA Response and across Endothelial Cell Subtypes . . . . .	67
3.3.7	Genetic Variation within VEGFA Responsive Sites . . . . .	68
3.3.8	Multiple Dynamic Changes in Transcription Factor Co-binding Upon VEGFA Stimulation . . . . .	69
3.4	Discussion . . . . .	74
3.5	Methods . . . . .	79
3.5.1	Data Generation . . . . .	79
3.5.2	ChIP-seq Data Processing, ChIP-seq Peak Calling for Transcription Factors and Co-activators . . . . .	79
3.5.3	RNA-seq Data Processing and Gene Expression Quantification and Defining VEGFA Responsive Genes . . . . .	80
3.5.4	GO BP Enrichment Analysis for Responsive Genes . . . . .	80
3.5.5	Chromatin State Calls using Hidden Markov Model Segmentation . . . . .	80
3.5.6	Enhancer Analysis . . . . .	81
3.5.7	DNase I hypersensitivity analysis . . . . .	81
3.5.8	Co-clustering of Transcription Factors . . . . .	81
<b>4</b>	<b>A Comprehensive Analysis of RNA Polymerase II Pausing Across Mammalian Cell Types</b>	<b>83</b>
4.1	Abstract . . . . .	83
4.2	Introduction . . . . .	84
4.3	Results . . . . .	84
4.3.1	Characterization of Pol II pausing across multiple cell types . . . . .	84

4.3.2	Pol II Pausing Influence On Gene Expression Levels . . . . .	88
4.3.3	Pol II Pausing and Cell Population Gene Expression Variability . . .	91
4.3.4	High Pol II TSS density promotes pausing release . . . . .	91
4.3.5	Many stimulus-responsive genes are paused and have lower PI prior to stimulation . . . . .	95
4.3.6	Pausing release selectively regulates rapid, signal-induced gene ex- pression change . . . . .	98
4.3.7	Pol II pausing relationship to the local chromatin landscape . . . .	100
4.3.8	Pol II pausing's relationship to chromatin topology . . . . .	105
4.4	Discussion . . . . .	106
4.5	Additional Methods . . . . .	108
4.5.1	ChIP-seq and RNA-seq analysis . . . . .	108
4.5.2	Calculating Pol II Pausing . . . . .	108
4.5.3	GO Analysis for paused genes . . . . .	108
4.5.4	Estimated mean PI coefficient of variation for genes within Hi-C TADs and ChIA-PET interactions . . . . .	109
<b>5</b>	<b>Conclusions and Future Directions</b>	<b>110</b>
5.1	Dynamic Chromatin Changes at Signal-Responsive Regulatory Elements . . . . .	110
5.2	p300 has Broad Effects on VEFGA-responsive Chromatin Remodeling . . . . .	112
5.3	Stimulus-Responsive Genes Have Structures Relating to Dynamic Gene Expression Change . . . . .	112
5.4	The Regulation of H2A.Z Deposition by Promoter-Proximal Pol II Pausing . . . . .	113
5.5	Implications for Understanding the Regulation of Angiogenesis . . . . .	114
<b>A</b>	<b>GWAS SNPs Mapping to HUVEC Enhancers and H3K27ac Variant Sites</b>	<b>116</b>
<b>B</b>	<b>Additional Tables and Figures For Promoter-Proximal Pol II Pausing Analysis</b>	<b>125</b>
B.1	Tables . . . . .	125
B.2	Figures . . . . .	126

# List of Figures

1-1	A crystal structure of a nucleosome. . . . .	15
1-2	A graphical summary of a typical CHIP-seq experiment workflow. . . . .	18
2-1	The experimental design of stimulating HUVECs with VEGFA . . . . .	31
2-2	Correlation of H3K27ac profiles from time series with ENCODE HUVEC data. . . . .	32
2-3	VEGFA-responsive H3K27ac change occurs near p300 binding sites in HUVEC. . . . .	33
2-4	Histogram of H3K27ac variance scores for windows with a minimum score of 4. . . . .	34
2-5	H3K27ac variant sites are much closer on average to p300 binding sites than less variant p300 sites. . . . .	34
2-6	H3K27ac profile with and without C646 treatment to compare changes in H3K27ac upon blocking p300 acetyltransferase activity. . . . .	35
2-7	H3K27ac variability decreased in C646 pre-treated cells. . . . .	36
2-8	Temporal scaling of the most H3K27ac variant windows near a p300 peak highlights three primary temporal patterns of H3K27ac variation. . . . .	37
2-9	Clustering the H3K27ac variant sites reveal three major temporal patterns in response to VEGFA. . . . .	38
2-10	Browser snapshot of example sites showing no significant change in DHS signal upon VEGFA stimulation at H3K27ac dynamic sites . . . . .	39
2-11	GREAT analysis of GO functional enrichments at variant H3K27ac sites. . . . .	40
2-12	ETS1 transcription factor strongly occupies p300 binding sites. . . . .	40
2-13	Transcription factor motifs enriched at H3K27ac variant sites. . . . .	41
2-14	All H3K27ac variant sites had an open chromatin structure at baseline and throughout the time course. . . . .	41
2-15	Relationship to VEGFA differential gene expression changes to H3K27ac variant sites. . . . .	43
2-16	VEGFA induces changes in chromatin looping frequency, while addition of C646 inhibits observed VEGFA responsive changes. . . . .	45
2-17	C646 pre-treatment of HUVECs inhibited VEGFA responsive gene change. . . . .	45
3-1	Clustering of VEGFA-responsive genes during the time course in order to identify temporal patterns of gene expression change during angiogenesis. . . . .	53
3-2	Protein coding and non-coding RNA gene count of each gene set. . . . .	54

3-3	Boxplot of expression ranges (in FPKM) for each of the four gene sets per time point. . . . .	55
3-4	Expression fold change relative to Hour 0 over time at VEGFA-responsive and non-responsive expressed genes. . . . .	56
3-5	Enrichment of GO biological process terms for each VEGFA responsive gene set. . . . .	58
3-6	Comparison of the distribution of gene lengths for each gene set. . . . .	59
3-7	Comparison of the distribution of transcription lengths for each gene set. . . . .	60
3-8	Unique exon count for VEGFA responsive gene sets. . . . .	60
3-9	Temporal dynamics at TSS of H3K27ac, H3K4me3, and DNase I hypersensitivity at VEGFA responsive genes. . . . .	61
3-10	Chromatin state model for VEGFA-stimulated HUVECs. . . . .	62
3-11	Distribution of chromatin states across HUVEC time course and ENCODE data. . . . .	63
3-12	VEGFA-responsive latent enhancers in HUVEC. . . . .	64
3-13	Several enhancer chromatin states lie within larger regions of repressed chromatin. . . . .	65
3-14	p300 binding frequency over time at super, latent, and island enhancers. . . . .	65
3-15	The changes in H3K27ac at super, latent and island enhancers over time with and without C646 treatment. . . . .	66
3-16	Number of long-range interactions to VEGFA responsive genes for super, latent, and island enhancers. . . . .	67
3-17	The distribution of H3K27ac variant sites across enhancer types and chromatin states. . . . .	68
3-18	Comparison of DNase I hypersensitivity peaks at HUVEC enhancers within the VEGFA time course. . . . .	69
3-19	DNase I hypersensitivity peak clustering across endothelial cell types and GM12878. . . . .	70
3-20	Summary of the transcription factor binding changes over time in response to VEGFA. . . . .	72
3-21	Average occupancy of ETS1 binding at SOM clustered candidate <i>cis</i> regulatory regions. . . . .	73
3-22	Average p300 occupancy over all time points across clustered candidate <i>cis</i> regulatory regions. . . . .	73
3-23	The average binding profile of ERG, FLI, and JUN occupancy. . . . .	75
3-24	RBPJ binding dramatically increases over time in response to VEGFA. . . . .	76
3-25	GATA2 and MYC have temporal, dynamic co-occupancy in response to VEGFA. . . . .	77
3-26	Frequency of candidate <i>cis</i> regulatory regions to nearest ERGs and LRGs. . . . .	78
4-1	Graphical summary of calculating a gene's Pausing Index (PI) . . . . .	85
4-2	Distribution of paused genes across cell types in human and mouse. . . . .	86
4-3	Analysis of enriched GO biological process terms in paused genes across cell types in order to identify recurrent functions of paused genes. . . . .	87
4-4	Highly paused genes tended to be the target of recurrent somatic mutations. . . . .	87

4-5	Comparison of GC percentage and CpG density at highly and lowly paused genes. . . . .	88
4-6	Enrichment analysis of all DNA 6- and 5-mner sequenecees at promoters supports combination of high GC and CpG content associated with more paused Pol II at TSS. . . . .	89
4-7	Paused and non-paused genes have a similar expression profile even while paused genes are the predominantly expressed gene in a cell type. . . . .	90
4-8	Trend modeling between gene expression and PI. . . . .	91
4-9	Paused genes have a lower expression variability cell-to-cell within a population across cell types. . . . .	92
4-10	Mean expression levels across a cell population between paused and non-paused genes did not not significantly vary. . . . .	92
4-11	To analyze the nature of Pol II activity and the PI, a visualization technique was developd to plot each gene by its Pol II TSSR and gene body density. Then the PI can be visualized in the left to right diagnoal. Each point/gene can be colored relative to a third variable, such as gene expression. When plotting all genes in GM12878 as such . . . . .	93
4-12	The inflection point within the TSSR-gene body Pol II trend line occured across cell types and generally located where the most highly expressed genes were. . . . .	94
4-13	Enrichment biological process GO terms across cell types for genes past the inflection point. . . . .	94
4-14	Deposition of NELF, CDK9, and CCNT2 at TSS relative to the PI. . . . .	95
4-15	Inflection point within TSSR-gene body Pol II trend curve only downregulated by total but not selective P-TEFb inhibition. . . . .	96
4-16	Knockdown of pausing release through various inhibitors shows that Pol II TSS density increases. . . . .	97
4-17	VEGFA responsive gene sets for pausing release analysis. . . . .	98
4-18	Both VEGFA responsive and non-responsive genes are paused prior to stimulation. . . . .	99
4-19	Signal-stimulus responsive genes have lower average PI at baseline than gene non-responsive to signal-stimulus. . . . .	99
4-20	Changes in pausing release in response to signal-stimulus does not uniformly increase or appear to primarily regulate all stimulus-responsive genes.100	
4-21	Inhibition of pausing release via flavopiridol in VEGFA treated HUVECs dramatically inhibits gene upregulation at selected genes tested. . . . .	100
4-22	Relationship between the PI and nucleosome depletion at TSS . . . . .	101
4-23	Histone marker to PI correlation across cell types. . . . .	102
4-24	H2A.Z stongly increases with increaing PI while H3.3 is not. . . . .	102
4-25	H2A.Z density around TSS increases with increasing PI across cell types. .	103
4-26	The nucleosome density around TSS for paused genes in mES increases highest at the most paused genes. . . . .	104
4-27	siRNA knockdown of H2A.Z in MCF7 cells globally increases Pol II pausing.104	
4-28	Distribution of how Pol II density changes after H2A.Z knockdown. . . . .	104

4-29	Comparison of H2A.Z enrichment relative to H3 density at selected promoters using ChIP-qPCR. . . . .	105
4-30	Lower PI coefficient of variation for genes within the same topological domain. . . . .	106
4-31	Genes having the same long-range interactions withn an enhancer based on Pol II ChIA-PET data have more similar PI than genes interacting with different enhancers. . . . .	107
B-1	Measurement of Pausing Index (PI). . . . .	127
B-2	Robustness of Pol II pausing measurements and calculations. . . . .	128
B-3	Pausing across cell and tissue types. . . . .	129
B-4	Relationship between whole gene, TSSR, and gene body Pol II density and PI to gene expression for H1, K562, IMR90, HUVEC, and HEPG2 human cells. . . . .	130
B-5	Relationship of gene expression to TSSR and gene body RNAP2 density and to PI. . . . .	131
B-6	Knockdown of H2A.Z in MCF7 cells. . . . .	132

# List of Tables

1.1	List of chromatin marks analyzed in this thesis. . . . .	19
3.1	ChIP-seq samples analyzed in this study. . . . .	52
3.2	Frequency of enhancers, and different enhancer types, in HUVECs. . . . .	63
A.1	Count of GWAS SNPs to H3K27ac variant sites and all HUVEC enhancers. . . . .	116
B.1	Summary of human and mouse cell lines used in Chapter 4 . . . . .	125
B.1	Summary of human and mouse cell lines used in Chapter 4 . . . . .	126

# Chapter 1

## Introduction

New blood vessels are formed through a process known as angiogenesis[1, 2]. While the cellular biology of angiogenesis has been extensively studied[1–4], the transcriptional regulation that underlies angiogenesis is still poorly understood[5]. With increased understanding of the non-coding genome[6–8] as well as newer methods to assay important, angiogenesis-related proteins genome-wide[9, 10], it is possible to observe angiogenesis-driven changes in histone modifications and transcription factors in order to identify candidate regulatory elements driving the changes. This will help both identify what regions of the genome regulate new blood vessel formation and lead to better understanding the transcriptional mechanisms that drive the process, in an effort to uncover novel therapeutic targets that can modulate pathological angiogenesis in human disease.

### 1.1 The Composition of the Human Genome and the Types of Genes

The human genome contains approximately 3 billion base pairs of DNA spanning 22 autosomal and 2 sex chromosomes[6], where a gene is the basic functional unit of a genome. Classically, genes encoded proteins[7, 11]. The human genome contains many genes that encode proteins, known as the *protein-coding genome*, and the remaining fraction, known as *non-coding genome*[6, 7]. Prior to the sequencing of the human genome, it was hypothesized that the genome encoded a large set of protein-coding genes to create the wide diversity of cellular phenotypes seen at the tissue level[12]. However after sequencing the human genome, only about 2% of the human genome, or 30,000 genes, appeared to be protein-coding[6]. Further analysis revised this initial estimate downward to around 24,000 protein-coding genes[13]. Overall, these estimates suggested that most of the human genome is non-coding. Yet, surprisingly, much of the non-coding genome was transcribed into RNA[6], which suggested functional activity. A substantial portion of this non-coding region is believed to contain regulatory elements, such as enhancers and promoters, that regulate cell-specific gene expression[14, 15], but the fraction is functionally important is still under debate. The ENCODE consortium estimated that up to 80% of the genome could be assigned biochemical activity from integrated analysis of transcription, epigenetic, and other genome-wide data sets[16], but only about 8.2% of the genome is



constrained over evolution[17, 18], which is a more traditional measure of whether a region of DNA is functionally important for cellular function. Hence, what fraction of the non-coding genome is functionally important requires further study.

Most protein-coding genes structurally contain a series of exons and introns with a promoter and transcription start site (TSS) at the 5' end of the gene and transcription termination site (TES) at the 3' end of the gene[19]. While a RNA polymerase transcribes a gene, the introns of that gene are removed from the nascent RNA by RNA splicing proteins co-transcriptionally[20–22]. For protein coding genes, the spliced RNA transcript is known as messenger RNA (mRNA)[19]. After polyadenylation, mRNA is moved to a ribosome in the cytosol in order to be translated into a protein[19]. An exon may be differentially included within a mRNA through *alternative splicing*, where the nuclear RNA splicing proteins determine the inclusion of a particular exon[19, 20]. Over 95% of known human protein-coding genes have more than one *isoform*, where the isoforms of a gene are the different RNA transcripts that are produced using different combinations of a gene's exons[20, 23, 24]. Different isoforms from the same gene may enable a gene to generate several different proteins from the same locus[19, 20], which contributes towards phenotypic diversity between cell types.

There are multiple types of non-coding RNA genes in addition to protein-coding genes[7, 25, 26]. Several well-known types are ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), and small nuclear RNAs (snRNAs)[7, 26]. These non-coding RNAs are important in the regulation of transcription and translation[26]. Over the past couple of decades, microRNAs (miRNAs) emerged as non-coding RNAs that bind and downregulate target RNAs through the Argonaute pathway[26, 27], similar to small interfering RNAs (siRNAs). A new major type of non-coding RNA of recent interest is long, non-coding RNA (lncRNA)[25, 28]. lncRNAs are similar to mRNAs[24, 28], such as having multiple exons and undergoing RNA splicing. However, lncRNAs do not generate any proteins and are more weakly conserved than protein coding genes[25, 28]. *XIST* and *H19* are a few classical examples of critical lncRNAs necessary normal cellular function[25]. Despite not generating any proteins, the knockdown of lncRNAs can have a negative effect on the cellular phenotype[25, 28, 29], suggesting that at least some lncRNAs are functionally important. But at this time, the function of most lncRNAs is still poorly understood.

## 1.2 The Organization of Chromatin and the Nucleus

Within the nucleus, the *chromatin*, a complex macromolecule consisting of DNA, RNA and protein[30], is packaged in an intricate three-dimensional structure[31, 32]. Chromatin has multiple functions[30, 33], including packaging the DNA and contributing towards the regulation of gene expression. During interphase, the chromatin is broadly divided into euchromatic and heterochromatic regions[34, 35], experimentally identified by the bands visualized during chromatin staining[35]. Euchromatin is chromatin with greater accessibility to other proteins and tends to contain actively transcribed genes[35]. Conversely, heterochromatin is chromatin with a more compact, less accessible structure that generally contains repetitive elements and repressed genes[34, 35]. Heterochromatin is further divided into two subtypes: facultative and constitutive[35]. Facultative heterochromatin is a

region that is repressed in a cell-type specific manner. Facultative heterochromatin arises through development, repressing genes that, for example, are of developmental importance but no longer need to be expressed in a particular cellular lineage[35]. Comparatively, constitutive heterochromatin is a region that is repressed in virtually all cell types[35]. For example, many regions of constitutive heterochromatin are composed of repetitive elements[35], such as the centromere. From these broad categories of chromatin, advances in molecular biology have allowed for a better understanding the molecular underpinnings of these regions.

At the molecular level, the basic unit of chromatin is the nucleosome[30]. The core of a nucleosome is made of DNA wrapped around a histone octamer, where histones are positively charged proteins that strongly interact with negatively-charged DNA[30, 36]. A major function of histones is compacting the nuclear DNA to fit within the microscale volume of the nucleus, since end-to-end the human genome spans a distance of over 1.8 meters[30]. There are five histone types: H1, H2A, H2B, H3, and H4[30, 36]. The H2A, H2B, H3, and H4 histones are known as the *core histones* because together these histones form the histone octamer of a nucleosome[30]. Each histone type has two protein copies within the histone octamer and has multiple genes that encode a histone protein that can incorporate into the appropriate position within the octamer[30, 36]. Folded histones have several basic amino acids on the protein's surface[30, 36], allowing the acidic DNA to wrap around the histone octamer[30], as seen in Figure 1-1. The core nucleosome has about 146 base pairs (bp) of DNA wrapping around the histone octamer[30]. The fifth histone type, H1, is known as the *linker histone* since it binds DNA between histone octamers. A nucleosome also contains *linker DNA* covered by H1, causing the amount of total nucleosomal DNA to vary[30, 35]. (The function of H1 is further reviewed in Harshman *et al*[37].) On average, the total nucleosomal DNA around the H1 histone and histone octamer is about 200 bp[30, 35]. Together the histones within a nucleosome mediate chromatin compaction and accessibility to the nucleosomal DNA. In euchromatic regions, nearby nucleosomes tend to be less compact, allowing greater openness or looseness of the DNA for increased accessibility by other DNA-binding proteins[30]. Heterochromatic regions tend to be more compact with nucleosomes folding into higher order structures, reducing accessibility of nucleosomal DNA[30]. While some higher order chromatin structures are well studied (*e.g.*, chromosome condensation during metaphase)[30], the nature of chromatin compaction is still under active study.

Histones were initially discovered by Albrecht Kossel, who biochemically isolated these proteins over a century ago[40]. While initially believed to be one protein, the different types of histones were uncovered over the next several decades[36]. Histones are highly conserved throughout the eukaryotic domain with low genetic variation across species[36]. The genes for the standard core histones are found in multiple copies in many eukaryotic genomes[36]. Each core histone also has multiple *histone variants* throughout the eukaryotic domain[36]. Compared to the genes encoding the standard core histones, a histone variant often has only a single gene encoding the variant protein[36], but these variants maintain a similar structure to its related standard core histone structure differing often only by a few amino acids and are often strongly conserved [36]. For example, H2A.Z is a very common histone H2A variant that is thought to regulate nucleosome stability[36, 41–43]. Some histone variants have been adapted to regulate specific functions[36]. For

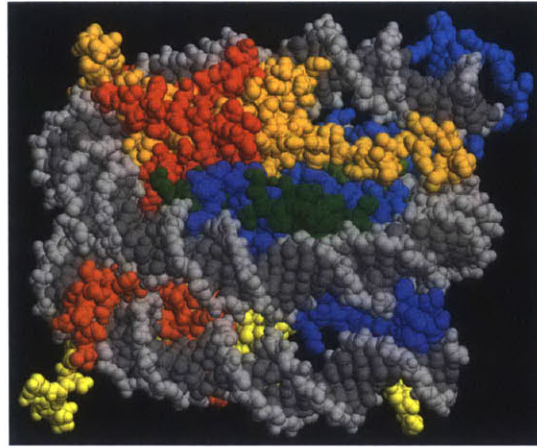


Figure 1-1: A crystal structure of a nucleosome with the nucleosomal DNA (gray) wrapped around the core histone octamer. On average, 146 bp of DNA wraps around the histone octamer. The core histone complex has 2 copies of each of the core histones (only half the nucleosome complex is visible here). Each histone within the crystal structure is colored as follows: H2A, yellow; H2B, red; H3, blue; H4, green. The raw crystal structure image can be found in the Protein Data Bank[38] under accession identifier 1EQZ[39].

example, CENP-A is specifically incorporated into nucleosomes in the centromere to aid centromere targeting during mitosis[36]. With all the histone variants across eukarya, each core histone type has a large gene family[36]. While the diversity of histones can allow for a variety of functional specializations, post-translational modifications to histones contribute towards regulating chromatin structure and activity.

There are a variety of post-translational modifications of histones that are linked with various cellular regulatory functions[8, 33, 44]. Prior to the identification of these post-translational modifications, histones were thought to be inert packaging proteins[36]. The initial discovery that suggested histones are involved in other regulatory roles was from Vincent G. Allfery and colleagues at the Rockefeller University in the 1960s[45]. He and his team discovered that individual histone fractions extracted from calf thymus could have a high density of acetylated lysines[45]. Acetylation of a histone would reduce its positive charge leading to repelling the DNA from wrapping around the histone, increasing chromatin accessibility and reducing chromatin packing. This led them to hypothesize that acetylating histones over a gene may increase its expression by increasing its chromatin accessibility for transcription[45].

The role of histones beyond DNA compacting was further expanded in 1988 with work by Michael Grunstein and colleagues on histone H4 in yeast. The N-terminal tail of each of the histones, like most of the rest of the histone protein itself, was strongly conserved throughout evolution[46], but the function of histone tails was not understood at the time. By mutating select amino acid residues along the H4 N-terminal tail in yeast, Grunstein and colleagues showed that mutations in the H4 tail caused the activation of normally silenced genes, although surprisingly it did not affect normally expressed genes or the yeast's growth[46]. The mutations also caused the cell cycle to lengthen[46]. Hence, this study

suggested that the N-terminal tail contributed towards normal cell cycle regulation in yeast through likely post-translational modifications, and that histones have other roles in addition to DNA compaction.

The N-terminal tail of each histone has multiple post-translational modifications identified[33, 44, 47]. Several amino acid residues on the N-terminal tail of a histone can be modified, such as lysines and arginines[44]. Many well-studied post-translational modifications are methylations and acetylations of these residues, but there are other known post-translational modifications, such as phosphorylation[33, 44, 47]. The expanding number of known post-translational modifications led to the development of the *histone code hypothesis*[48]. This posited that the combination post-translational modifications on the N-terminal tail could regulate protein-binding to a specific region instead of only affecting the DNA-histone interaction[48], since the expanding number of known post-translational modifications allowed for a large combinatorial code that could be used for regulating gene expression and other functions. But it also appears many post-translational modifications are functionally redundant[33, 49, 50], narrowing the potential number of distinct combinations of post-translational modifications and the span of the histone code.

Another question that arose with the expanding number of identified histone post-translational modifications is what enzymes add, remove and read these post-translational modifications[30, 44]. Each post-translational modification to the N-terminal tail of a histone has a corresponding *writer* and likely *eraser* enzyme[30, 44]. A writer or eraser enzyme can modify one or more amino acid residues on a N-terminal histone tail. For example, p300 and CREB binding protein (CBP) both contain acetyltransferase domains that acetylate lysines on histone H3[44, 51]. In addition to reading and writing, many proteins contain chromatin binding domains that act as *readers* for one or more N-terminal histone tail post-translational modifications[44]. A notable example is heterochromatin protein 1 (HP1) that recognizes and binds H3K9me3 that then condenses nearby chromatin when bound[33, 44]. Although these modifications have no likely catalytic activity themselves, the establishment and removal of post-translational modifications still can contribute towards cellular regulation through altering the chromatin structure and the localization of chromatin-binding proteins.

In order to better understand the organization of chromatin, multiple methods have been developed to assess occupancy of chromatin bound proteins[9, 52–55]. One key method to assess the chromatin occupancy of a target protein is chromatin immunoprecipitation (ChIP)[52]. ChIP is the immunoprecipitation of a target protein, using an antibody raised against the target protein, with the associated chromatin fragment in the co-immunoprecipitate[52]. In order to obtain chromatin fragments from cells, the cells are lysed, and the chromatin is fragmented into small segments (such as through sonication). Often a chemical agent, such as formaldehyde, is added prior to lysing the cell to cross-link proteins and DNA in chromatin to improve the likelihood of co-immunoprecipitating DNA during the ChIP pulldown. Then the enrichment of a target protein at a selected DNA region can be assessed through quantitative polymerase chain reaction (qPCR). ChIP was first developed to assay the position of RNA Polymerase II (Pol II) at target sites in the *Drosophila* genome during studies about transcription[56]. However, a major limitation of ChIP-qPCR is that it only assesses one region at a time.

Effective parallel assessment of many chromatin regions via ChIP was enabled with the

development of microarray technology in the 1990s[57]. Although initially developed to assay the expression of multiple genes in parallel[57], microarray technology was adapted for ChIP experiments (known as ChIP-chip or ChIP-on-chip)[55]. Microarrays are platforms that contain multiple DNA probes at selected genomic regions of interest[55, 57]. After marking the DNA pulled down from a ChIP with a fluorescent tag, this extracted DNA is put into the microarray such that the tagged DNA hybridizes with the probe sequences on the chip[55]. Probe regions enriched with the target protein will be increasingly hybridized by tagged DNA from the ChIP pulldown, which can be measured by the fluorescence intensity at a specific probe[55]. This technology was successfully applied to multiple histone modifications and other chromatin proteins across multiple organisms[49, 58]. Yet, especially for large genomes like humans, this did not cover the entire genome due to limitations of the number of probes that could be placed on a microarray[10, 44, 55], requiring newer technologies for true genome-wide analysis of histone modifications and other chromatin bound proteins.

The development of high-throughput sequencing led to the development of ChIP followed by high-throughput sequencing[9] (ChIP-seq). Instead of hybridizing the recovered DNA sequences from a ChIP to a microarray, the extracted DNA sequences from the ChIP pulldown are sequenced and mapped back to the target genome to assess the position of the targeted protein[9, 10]. (A graphical summary of a ChIP-seq experiment is provided in Figure 1-2.) ChIP-seq can provide base pair resolution of the position of a target protein, although there are many technical challenges associated with analyzing ChIP-seq data[10]. In order to address some of the experimental challenges, alternative methods have been developed, notably DamID and ChIP-exo[53, 54]. DamID allows assaying a chromatin-bound protein without cross-linking agents or antibodies by affixing an adenosine methyltransferase to the target protein[54]. The attached adenosine methyltransferase then methylates adenosines at GATC sites when the target protein is bound to chromatin, and then the methylation status can be assessed with a methylation sensitive restriction enzyme [54]. However, DamID has lower resolution than ChIP-seq (since a GATC sequence occurs once every 1024 bp on average) and requires creating a fusion protein[54]. ChIP-exo, on the other hand, uses an antibody pulldown like ChIP-seq but improves the data profile resolution by first digesting the extracted chromatin DNA with an endonuclease to cut the DNA fragments down to the protected region by the target protein before completing the rest of ChIP-seq[53], reducing the noise in the experiment by only sequencing DNA fragments right around the target protein.

The analysis of ChIP-seq data first requires aligning sequenced DNA fragments back to a reference genome[9, 10], such as the human genome. There are several programs to align high-throughput sequencing data, such as *bowtie*[60] and *bwa*[61]. The alignment of these DNA reads assesses the position of the target histone modification or other chromatin protein along the genome. In addition to the ChIP sample, an input DNA profile is often generated to account and correct for some ChIP-seq related biases in the signal profile[9, 10]. Using ChIP and input data sets, ChIP-seq peaks can be called using tools such as MACS[62] and *spp*[63]. ChIP-seq peaks are regions of significant enrichment in the ChIP-seq profile that reflect likely positions where the target protein is located.

Several ChIP-chip and ChIP-seq studies over the past decade across metazoan organisms mapped post-translational histone modifications genome-wide in order to better un-



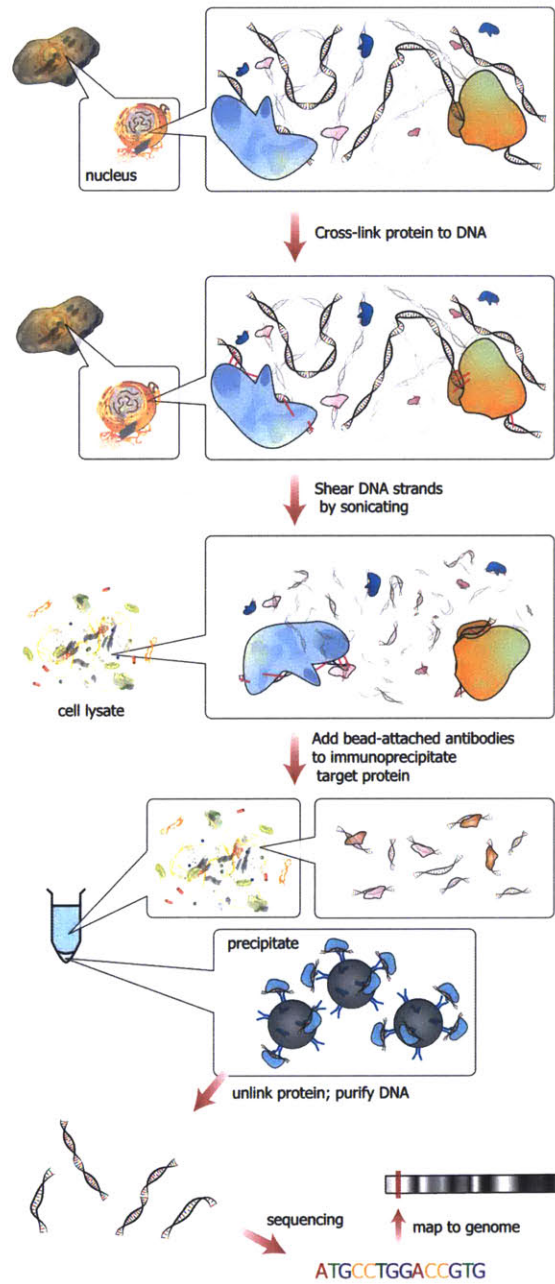


Figure 1-2: A graphical summary of a typical ChIP-seq experiment workflow[9]. (Image licensed and used under CC BY-SA 3.0[59].)

<b>Chromatin Mark</b>	<b>Brief Summary</b>	<b>References</b>
H3K4me3	Marks TSSs of expressed genes	[8, 65]
H3K4me1	Marks enhancers	[8, 15, 65]
H3K4me2	Marks enhancers	[8, 66]
H3K27ac	Marks active TSSs and enhancers	[8, 15, 58, 67]
H3K27me3	Marks repressed regions (facultative heterochromatin) through PRC2	[8, 68]
H3K9me3	Marks repressed regions (constitutive heterochromatin)	[8, 44]
H3K36me3	Marks regions recently transcribed by Pol II, usually at the gene of gene bodies over exons	[8, 69]
H2A.Z	Marks promoters (both active and sometimes inactive) and active enhancers	[8, 70, 71]
DNase I Hypersensitivity	Identifies regions of open chromatin, such as at TSSs and active enhancers	[72, 73]

Table 1.1: List of chromatin markers analyzed in this thesis. DNase I hypersensitivity is not a chromatin mark, *per se*. But it is included here because of its utility in identifying regions of open chromatin.

derstand the organization of the chromatin[8, 58, 64, 65]. A brief summary of the commonly attributed function(s) of each histone modification studied in this thesis is listed in Table 1.1, and a more detailed summary of these histone modifications is provided over the next several paragraphs.

H3K4me3 (tri-methylation of lysine 4 on histone H3) is primarily enriched at the TSS of expressed genes[8, 58, 64, 65, 68] (see Table 1.1). This mark is established likely through the binding of Pol II at a promoter and subsequent recruitment of the Trithorax complex to establish it[74]. Consequently, repressed genes typically lack H3K4me3 enrichment[8]. Since Pol II initiation and H3K4me3 enrichment at TSS is linked, the ChIP-seq signal intensity of H3K4me3 at TSS correlates with a gene's expression level[64, 75]. While H3K4me3 enrichment has been often studied at protein coding genes, expressed non-coding RNA genes also have H3K4me3 enrichment. This includes lncRNAs[28] and transcribed repetitive elements, such as tRNAs[76], suggesting H3K4me3 establishment is independent of the whether a particular gene encodes a protein.

One potential function for H3K4me3 is to maintain an open chromatin structure at

TSS[74], where an open chromatin structure reduces the barrier for Pol II initiation[77]. In addition to maintaining an open chromatin structure at a promoter, H3K4me3 establishment may positively influence Pol II initiation as well. For example, at least one of the Pol II general transcription factors binds H3K4me3 to help recruit the pre-initiation complex[78]. In a complementary report, broad, strong enrichment of H3K4me3 at genes can stabilize Pol II to reduce their gene expression variability across a population of cells[79].

Unlike at TSS, H3K4me3 is generally absent at distal regulatory elements[8, 58]. However, sometimes distal regulatory elements may have H3K4me3 enrichment, such as those that are actively transcribed in humans[24] and mouse[80]. This was consistent with other reports that Pol II bound and transcribed a subset of active enhancers[81, 82]. Hence, Pol II binding and transcription at mammalian enhancers may induce the establishment of H3K4me3, but whether these H3K4me3-enriched distal regulatory elements are actually independent enhancers or part of a yet unidentified gene[83] remains to be further studied.

H3K4me1 (mono-methylation of lysine 4 of histone H3) tends to mark cell-specific enhancers within the chromatin (see Table 1.1)[15, 65]. Unlike H3K4me3, H3K4me1 is typically enriched at distal enhancers rather than at TSS across species[58, 65, 82], although H3K4me1 is weakly enriched relative to H3K4me3 at the TSS of expressed genes[64]. But the presence of H3K4me1 does not guarantee that an enhancer is actively regulating transcription[82]. Generally, active enhancers are better identified by the combination of H3K4me1 and H3K27ac (described further below)[82]. Hence, H3K4me1 appears insufficient to activate an enhancer, but it is useful in identifying candidate enhancers within a cell type[82]. Another histone methylation, H3K4me2 (di-methylation of lysine 4 on histone H3), tends to co-occupy regions of H3K4me1[8, 84], and it has been used to identify sites undergoing dynamic changes in nucleosome positioning in response to external stimuli[66]. H3K4me2 is also found with H3K4me1 and H3K4me3 at an expressed TSS[8, 64], but H3K4me3 is still more strongly enriched with increasing expression than H3K4me2[64].

Recently, it has been shown that H3K4me1 may be established *de novo* in response to external stimuli[85] (in combination with H3K27ac). This *de novo* H3K4me1 enables inactive enhancers that, at least temporarily, create cellular memory of recent regulatory activity because, in macrophages, latent enhancers reactivate rapidly when re-stimulated by the same initial stimulus that created the latent enhancer[85]. Although H3K4me1 may not cause an enhancer to be active, the establishment of H3K4me1 appears important for normal enhancer function.

H3K27ac (the acetylation of lysine 27 on histone H3) marks both active TSSs and enhancers throughout the genome (see Table 1.1)[8]. When a histone is acetylated, it influences transcription and chromatin accessibility by repelling the nucleosomal DNA from the core histone octamer[30, 45]. At an expressed TSS, H3K27ac tends to be found in combination with H3K4me3[8], and its enrichment also scales with a gene's expression level[75]. At enhancers, the establishment of H3K27ac marks an active enhancers with a greater sensitivity than H3K4me1[8, 67, 82]. But similar to H3K4me1, *de novo* H3K27ac can be established in response to external stimuli to activate latent enhancers[85]. The establishment of H3K27ac is driven by the binding of transcription factors and recruitment of histone remodelers to the bound enhancer or promoter[86]. Two important examples of such chromatin remodelers are p300 and CREB binding protein (CBP), both of which have histone acetyltransferase domains that target H3K27[51, 87]. Given that H3K27ac



tends to mark active enhancers in a cell type, its establishment tends to mark cell-type specific enhancers[58, 67], making H3K27ac a useful mark to profile in order to identify such enhancers.

H2A.Z (also known as H2AFZ) is a H2A histone variant [36, 88]. Like other histones, H2A.Z is highly conserved across all eukaryotic taxa, and it is critical for normal function[36, 70]. H2A.Z is thought to help destabilize the nucleosome to increase nucleosome turnover and chromatin accessibility[41–43]. In mammals, H2A.Z tends to be found at many TSSs and active enhancers[14, 42, 64], and appears to be increasingly incorporated with higher expressed genes[42, 64, 89]. It also helps mediate the binding of transcription factors by opening chromatin and, possibly, helping recruit them directly[8, 14, 42]. Although strongly present at active genes, it also is deposited at Polycomb-repressed promoters to help recruit Polycomb Repressive Complex 2 (PRC2)[42]. Overall, H2A.Z incorporation is rather broad and requires further study to understand its full function.

H3K27me3 (tri-methylation of lysine 27 of histone H3) is a developmental repressive mark that is found at facultative heterochromatin[68, 90]. This mark tends to repress developmental regulators as the cell differentiates[8, 91, 92]. PRC2 establishes H3K27me3 through the use of EZH1 or EZH2 within the complex [8, 93]. Although generally thought to be repressive, H3K27me3 may be found in co-occupancy with H3K4me3 at TSS of genes that have *bivalent domains*[94], which are primarily found in embryonic stem cells[93] but also exist in differentiated cell types[94]. Bivalent domains tend to occur at genes regulating development and tend to have low expression despite the presence of H3K4me3[90, 93, 94]. During differentiation, many of these bivalent domains become either activated or repressed by keeping H3K4me3 or H3K27me3, respectively[93, 94]. However, some bivalent domains may remain after differentiation with low expression[93], although the role of bivalent domains in non-embryonic stem cells remains unclear.

A second major heterochromatic histone modification is H3K9me3 (tri-methylation of lysine 9 on histone H3)[64, 90]. This histone modification tends to mark constitutive heterochromatin[8, 64]. Many repetitive elements are repressed through H3K9me3 enrichment across cell types[76]. H3K9me3 can be established through miRNA-mediated chromatin remodeling as miRNAs silence the transcription of targeted regions[95]. H3K9me3 represses gene expression through recruiting HP1 that both compacts nearby chromatin and blocking Pol II[44]. In humans, H3K9me3 and H3K27me3 tend to be exclusively enriched[96], even though both mark heterochromatin.

H3K36me3 (tri-methylation of lysine 36 on histone H3) is a mark that covers gene bodies of expressed genes[8, 90]. In comparison to the H3K4 methylations and H3K27ac, it occurs downstream of a gene's TSS[90]. This mark is established through Pol II elongation[44], often being more enriched at the exons of a gene rather than its introns. The enrichment of H3K36me3 at exons helps mediate RNA splicing since several RNA splicing factors bind H3K36me3[97]. Another role for H3K36me3 may help stabilize nucleosomes since it suppresses nucleosome turnover in yeast in order to suppress aberrant transcription[44, 69], although it is unclear if this is still its function in other organisms. Additionally, H3K36me3 may regulate transcription indirectly. A recent report showed that H3K36me3 can be bound by ZMYND11[98]. ZMYND11 when bound inhibits transcription by blocking the release of Pol II from its promoter-proximal pausing state[98] (Pol II pausing is discussed in Section 1.6). Although H3K36me3 is established by Pol II

as a consequence of elongation, it has multiple roles in regulating gene activity.

Another informative property of chromatin structure is identifying regions of open chromatin may contain active regulatory elements[99]. Histones can occlude access to the nucleosomal DNA, which can inhibit the binding of transcription factors and other chromatin binding proteins[72, 73, 86, 99]. In order to improve binding by transcription factors and other chromatin proteins, chromatin needs to open and expose the underlying DNA for binding[14, 86]. Hence, identifying of open chromatin regions can be informative of active regulatory regions within a cell type, which can be achieved through DNA cleavage enzymes[72, 73, 99]. Under the proper experimental conditions, open chromatin regions are more readily digested by DNA cutting enzymes, such as DNase I[72, 99]. Regions preferentially cut by DNase I are known as *DNase I hypersensitive regions*. Promoters and active enhancers tend to be DNase I hypersensitive[73], but many promoters are DNase I hypersensitive across many cell types[73], suggesting that promoter accessibility contributes little to cell-type gene expression patterns. Comparatively, the enhancers tend to have cell-type specific DNase I hypersensitivity [72, 73], suggesting that the activity of distal regulatory elements is tightly controlled across cell types through altering chromatin accessibility[14].

### 1.3 Transcription Factors Bind Regulatory Elements and Regulate Transcription

Transcription factors are DNA binding proteins that bind an enhancer or gene promoter to regulate transcription at a target gene[100, 101]. Each transcription factor has a DNA binding domain and regulates the activity of a RNA polymerase when bound to chromatin[100, 101]. The human genome is estimated to contain a large number of potential transcription factors[100, 101], suggesting that the regulatory network in humans is rather diverse and complicated. There are two groups of transcription factors differentiated by their DNA binding properties: general or sequence-specific. General transcription factors make up RNA polymerases, such as Pol II, and bind gene promoters[102]. Sequence-specific transcription factors are the larger set of transcription factors that each have a *consensus sequence* bound by its DNA binding domain[100, 101]. The size of a consensus sequence ranges from small (such as the 4 bp core motif for the GATA family of transcription factors) to large (such as p53, see the JASPAR database for more examples[103]). The consensus sequence of a transcription factor can be identified by analyzing ChIP-seq data, in combination of consensus motif estimating software such as MEME[104], DREME[105], or MEME-CHIP[106], more sensitive techniques, such as SELEX[107] or protein binding microarrays[108]. Many transcription factors are grouped into families. For example, the ETS transcription factor family is a prevalent set of transcription factors expressed throughout many cell types and where all factors generally bind the core GGAA DNA motif (or a close variation)[109]. Often multiple ETS transcription factors are expressed within a human cell type, but how they achieve binding specificity from each other expressed ETS factor remains under study[109, 110].

Many transcription factors must co-bind with other transcription factors or co-activators to drive changes in gene expression[86]. Sometimes the binding of a single transcription

factor may be sufficient, possibly such as MYC over-expression in many cancers[111, 112]. But, especially in humans, the co-binding of transcription factors appears necessary for transcriptional change[14, 86]. Co-activators, such as p300 and CBP, help stabilize the co-binding of multiple transcription factors[14, 51, 86]. Additionally, the binding of multiple transcription factors is strongly influenced by the chromatin structure[14, 73, 86], possibly requiring a pioneer factor to open chromatin first before the co-binding of other factors[86, 113]. DNase I hypersensitivity analysis can identify such regions of open chromatin that transcription factors could bind, although a recent report suggests that co-binding transcription factors can bind stably at non-DNase I hypersensitive regions[114]. Overall, the co-binding of transcription factors is heavily influenced by both the DNA content and chromatin structure of enhancers in a cell type.

## 1.4 Enhancers Regulate Gene Expression by Looping Towards Target Promoters

Enhancers are DNA regulatory elements that regulate gene expression of a target gene or genes, often in a cell-type specific manner[14, 15, 115]. Enhancers regulate transcription by serving as platforms for transcription factor binding[14]. Unlike gene promoters, the chromatin structure of enhancers tends to be more cell-type specific to regulate its activity[14, 15, 50, 58, 73, 82]. Active enhancers tend to be marked by both H3K27ac and H3K4me1 in addition to being DNase I hypersensitive[14, 15, 82]. Additionally, H2A.Z and H3.3, a H3 histone variant[36], may occupy an active enhancer as a way to increase its nucleosome turnover and accessibility[14, 15]. Together, these changes in chromatin structure increase its accessibility to allow transcription factor binding.

The how transcription factors must bind to an enhancer to cause gene expression changes is still under study[86]. Especially in humans, enhancers often contain multiple transcription factor binding sites of different sequence-specific transcription factors, and multiple transcription factors may need to co-bind for transcriptional change[14, 86]. Pioneer factors are thought to bind first in order to open the chromatin structure of an enhancer to enable binding of other transcription factors[86, 115]. After this, a basic model would suggest all transcription factor binding sites need to be bound at an open enhancer to cause transcriptional change[86]. Yet, there is increasing evidence that not all transcription factor binding sites may need to be bound to be functional[86], leading to further questions on how to determine the necessary set of transcription factors at an enhancer for driving gene expression.

Recent studies have begun to classify different types of enhancers[85, 116, 117]. One popular class is *super enhancers*, which are enhancers with extremely strong H3K27ac and/or BRD4 signal[116, 117], where BRD4 is a transcriptional co-activator. Super enhancers may include multiple adjacent peaks of H3K27ac and/or BRD4[116, 117], allowing a super enhancers to span a large region. While the differences between super enhancers and normal enhancers are still under study, super enhancers are important given their general proximity to and ability to regulate key cell identity and developmental genes[116, 117], such as *OCT4* in embryonic stem cells. With better identification of candidate en-

hancers in a cell type through the use of profiling certain histone marks[15, 58, 82], candidate enhancer identification has becoming increasingly easier leading to the need to potentially further classify different types of enhancers to understand their function.

Enhancers are thought to regulate transcription by creating chromatin loops towards their target gene (or genes) after the binding of transcription factors[14, 118, 119]. Two key complexes are known to help form of chromatin loops: Mediator and cohesion[119–121]. These complexes are recruited to enhancers upon changes in transcription factor binding and drive looping towards the target promoter. Newer experimental methods have enabled improved analysis of enhancer-promoter and other chromatin loops genome-wide[32]. Hi-C[32] has been particularly useful in capturing this chromatin structure. Recent studies have shown the chromatin is foled into multiple active or repressed compartments.[31, 32]. Some of these compartments form topologically associating domains (TADs)[122, 123], where most enhancer-prmoter chromatin loops tend to occur within a TAD[122]. Hence, the location of TADs are important determinants in regulating enhancer targeting[14, 122, 123]. TADs tend to be also stable across cell types[122], suggesting variation in TADs may weakly contribute to gene regulation. But recent extremely high-resolution Hi-C data showed that there are a number of cell-type specific chromatin loops within larger compartments and TADs[124], suggesting that enhancer-promoter chromatin loops tend to be shorter and rather cell-type specific.

## **1.5 Chromatin Remodeling During Development Compared with Chromatin Remodeling During Signal Response**

Chromatin remodeling is a major component of cellular differentiation during development[90–92, 125–127]. The chromatin landscape remodels as a cell differentiates in order to cause changes in active regulatory elements that, in turn, will drive changes in gene expression to establish the new cell phenotype[90–92, 125, 127]. In embryonic stem cells, most of the genome begins in a euchromatic state[90, 128]. As a cell differentiates, the chromatin landscape is remodeled to alter the accessibility of genes and enhancers active within a particular cell type[128]. This also includes the expansion of heterochromatin as most differentiated cells tend to have a larger heterochromatin content than euchromatin[128, 129] and the remodeling of bivalent domains [90, 94].

In terminally differentiated cells, the importance of chromatin remodeling less well understood. Multiple studies have suggested that chromatin remodeling accompanies cellular response to external stimuli[66, 85, 130–133] or even normal rhythms[134]. For example, changes in nucleosome positioning may occur in response to an external stimulus through changes in transcription factor binding[66, 133], possibly converting heterochromatic regions into euchromatic regions. Chromatin remodeling has also been reported to occur at the TSS of genes expressed in response to external stimuli[132]. In neurons, changes in H3K27ac were observed at candidate neuronal enhancers in response to depolarization of the neuron[131]. In mouse dendritic cells, multiple extracellular stimuli cause the activation of latent enhancers, *de novo* H3K4me1 and H3K27ac deposition at previously inactive regions[85]. While there are examples of chromatin remodeling in terminally differenti-

ated cells, few studies have studied temporal changes of chromatin structure over short time courses and connected it with downstream changes in gene expression.

## **1.6 The regulation of Pol II and transcription by regulating both initiation and the entry into elongation**

RNA polymerases transcribe genes into RNA under many layers of regulation[77, 135, 136]. Animal genomes generally contain three RNA polymerases. (Comparatively, plants have five RNA polymerases[137, 138].) Each RNA polymerase transcribes a particular set of genes. RNA Polymerase I transcribes several rRNA genes[102]. RNA Polymerase III transcribes tRNA genes as well as some rRNA genes[102]. Pol II transcribes the majority of genes since it transcribes both protein coding and long, non-coding RNA genes, in addition to snRNA and miRNA genes[102, 136]. Pol II is a multi-protein complex that contains several of general transcription factors[102, 136]. After establishing the Pol II pre-initiation complex at a promoter[102], Pol II begins transcribing the gene; but, before entering productive elongation, Pol II pauses just downstream and proximal to the TSS[77, 135, 136]. This Pol II promoter-proximal pausing often occurs between 20 to 60 bp downstream of the TSS[77, 135, 139], and it remains paused until it is released to begin elongation. The pause is caused by the binding of negative elongation factor (NELF) and DRB-sensitive inducing factor (DSIF) to Pol II[77, 135]. Additionally at this point, the 5-serine on the C-terminal domain (CTD) of Pol II is also phosphorylated[77, 136]. Pol II remains in this paused state until released through the recruitment of positive transcription elongation factor (P-TEFb)[77, 135, 136]. P-TEFb removes NELF and phosphorylates serine 2 on the Pol II CTD tail in order for Pol II to begin elongation[77, 136]. The P-TEFb complex includes the 7SK lncRNA and CDK9[136], and the activity of P-TEFb is blocked by the small molecular inhibitor flavopiridol[135, 136, 140] (FP). After pausing release, elongating Pol II may pause elsewhere along the gene body (*e.g.*, during co-transcriptional splicing) even though these other pauses appear currently unrelated to the regulation of promoter-proximal Pol II pausing[135, 141].

This promoter-proximal pausing was initially discovered at *Drosophila* heat shock genes[77]. By heat shocking *Drosophila* cells, there was a surprising rapid transcriptional upregulation on a set of heat shock genes, faster than expected from the time it took to transcribe a non-heat shock gene[77]. Initially, it was hypothesized that Pol II pausing helped mediate rapid changes in gene expression since pausing would maintain stable, transcriptionally engaged Pol II for rapid activation[77, 142]. Since then, several studies showed that Pol II pausing was widely present at non-signal responsive genes in the human and *Drosophila* genomes[143–145], and Pol II pausing was not necessary for rapid changes in gene expression[142]. Since then, estimates of the number of paused genes range from 30% to 90% of all genes within a genome[77], depending on the technology and criterion used. But this suggests a widespread rather than narrow phenomenon in transcriptional regulation.

With increasing understanding of Pol II pausing, several reports have identified that transcriptional output can be regulated through altering pausing release rather than only

Pol II initiation[111, 116, 117, 146]. Traditionally it was believed that transcription factors regulated gene expression by influencing Pol II initiation[77]. However, transcription factors can recruit P-TEFb to drive pausing release and increased gene expression, such as *MYC*[111]. Enhancers, including super enhancers, also can drive increased pausing release instead of Pol II initiation[116, 117, 146].

Pol II pausing can be measured through a variety of different methods[77, 135]. The rate of Pol II pausing and release can be measured directly through *in vivo* methods that visually observe the stalling and release of Pol II at a single gene[77, 135]. Estimating Pol II pausing on a genome-wide scale currently involves inferring the paused state from an assay that measures Pol II occupancy and estimating the relative fraction of the Pol II at a gene's TSS versus Pol II in its gene body. Multiple genome-wide methods exist for measuring Pol II occupancy[77, 135, 139, 141, 144, 145]. ChIP-seq against Pol II maps the location of all chromatin-bound Pol II. However, ChIP-seq is insensitive to whether Pol II is transcriptionally engaged (*i.e.*, whether a Pol II complex on the chromatin is actively transcribing)[77, 135]. A Pol II ChIP-seq profile is also sensitive to the quality of the antibody used [77, 145]. For example, changes in the phosphorylation of the Pol II CTD region as Pol II elongates along the gene body may affect the binding of the antibody and the pulldown of Pol II[145]. Alternative methods have been developed to avoid limitations of ChIP experiments by pulling down the nascent transcript in Pol II (*i.e.*, the growing RNA transcript as Pol II elongates along a gene). These methods include GRO-seq[145], PRO-seq[139], and NET-seq[141]. GRO-seq[145] measures the of location of transcriptionally engaged Pol II complexes. PRO-seq[139] and NET-seq[141] are newer variants with a more precise mapping of the position of Pol II. All three methods quantify transcriptionally engaged Pol II better than Pol II ChIP-seq[135], although Pol II profiles from ChIP-seq Pol II have been shown to be almost all from engaged Pol II when compared against GRO-seq profiles[147].

## 1.7 The Physiological Process of Creating New Blood Vessels: Angiogenesis

Blood vessels are an essential part of the cardiovascular system as they perfuse almost every organ in the body. Phenotypically blood vessels are diverse in size and structure depending on their location in the body, but endothelial cells form the basis of all blood [1]. Endothelial cells help maintain the vessel integrity and regulate new blood vessel formation [1, 2, 148]. The initial vasculature tree is formed during early development[1, 2]. From the inner cells of the embryo, the hematopoietic progenitor cells begin to form in blood deposit regions. From these hematopoietic progenitors, the endothelial progenitor cells (EPCs) differentiate[1]. EPCs then further differentiate into endothelial cells that in turn create the early vasculature. After embryogenesis, EPCs were thought to not be further involved in the formation of new blood vessels. However, recent studies suggest that EPCs may be still active in blood vessel formation post-development[1].

In adults, new blood vessels form from existing vessels, a process known as *sprouting angiogenesis*[4, 149]. Sprouting angiogenesis is triggered by stimulation from the binding

of an angiogenic signal, such as vascular endothelial growth factors (VEGFs), at a complementary surface receptor on an endothelial cell, such as vascular endothelial growth factor receptors (VEGFRs). A VEGF is secreted by a tissue when it requires increased vascularization to meet its metabolic demand. The stimulated VEGFRs on endothelial cells activate transcriptional changes that drive angiogenesis[4, 150]. This causes some endothelial cells to become *tip cells* and the rest *stalk cells*[3, 4]. Tip cells lead the formation of new blood vessels by migrating towards the source of the secreted factor, and stalk cells follow migrating tip cells[3, 4]. The determination of which endothelial cells are tip versus stalk cells appears to be mediated through intercellular signaling via the Notch pathway[3, 4]. The endothelial cells receiving the strongest VEGF signal become tip cells and then signal to adjacent endothelial cells to become stalk cells by driving gene expression changes in these stalk cells. As the endothelial cells migrate, they must remodel the surrounding matrix while forming the lumen of the new blood vessel until reaching the source to complete the process[4, 149].

Angiogenesis is a critical component of a variety of human diseases[1, 151]. One major area of study is tumor angiogenesis[151, 152]. As a tumor grows, it requires increased vascularization for its increased metabolic demand and growth[151, 152]. Tumors secrete VEGF and related angiogenesis stimulating factors into the local tissue environment, and blocking this signaling negatively impacts tumor growth[151, 153]. Aside from tumor angiogenesis, diabetic retinopathy, the abnormal growth of blood vessels in the retina that progressively leads to blindness, is a very common complication of advanced diabetes[1]. Additionally, angiogenesis has been increasingly implicated in a variety of human diseases, such as stroke, Alzheimer's Disease, and kidney nephropathy[1].

Given the role of angiogenesis in several diseases, therapeutically controlling angiogenesis is a major area of study in drug development[1, 151, 153]. A major anti-angiogenic drug on the market is bevacizumab (whose trade name is Avastin)[154, 155]. Bevacizumab is a recombinant, monoclonal antibody that targets VEGFA. Its use has been approved in the United States for treatment of multiple types of cancer as well as certain diseases involving abnormal angiogenesis[154, 155]. Aside from therapeutically inhibiting angiogenesis, stimulation of angiogenesis through VEGF is under study as a treatment for coronary artery disease[156]. In coronary artery disease, the blood flow within coronary arteries becomes progressively blocked, leading to ischemia and heart failure. It is thought that increasing VEGF in the heart may improve coronary heart disease by increasing vascularization in order to bypass blockages[156]. Hence, an improved understanding of the regulation of angiogenesis may lead to novel drug targets to regulate angiogenesis.

## **1.8 Understanding the Chromatin Dynamics and Transcriptional Regulation of Angiogenesis**

By analyzing changes in gene expression, chromatin structure, enhancer activity and transcription factor binding changes during angiogenesis, this will uncover the temporal regulation of the process. In Chapter 2 of this thesis, the dynamic remodeling of H3K27ac in response to VEGFA identified candidate regulatory elements linked with angiogenesis.

These candidate regulatory elements had multiple temporal patterns of H3K27ac dynamics linked with corresponding changes in differentially expressed genes and chromatin looping. Additionally, p300 was found to be necessary in angiogenesis to cause changes in H3K27ac enrichment and chromatin looping. In Chapter 3, a broader analysis gene expression, chromatin and transcription factors during angiogenesis allowed for the dissection of the temporal changes in transcriptional regulation. Multiple non-coding RNA genes were found to be co-expressed with differentially expressed protein-coding genes, and multiple VEGFA-induced transcription factor co-binding patterns were identified linked with many of the genes responsive to VEGFA. In Chapter 4, patterns of Pol II promoter-proximal pausing in relationship to gene expression and chromatin structure were studied across many mammalian cell types. In angiogenesis, this Pol II pausing was present at most VEGFA-responsive genes but only appeared to drive responsive gene rapidly regulated to VEGFA. Additionally, multiple recurrent patterns across cell types of Pol II pausing to gene expression and chromatin structure were identified.



## Chapter 2

# A dynamic H3K27ac signature identifies VEGFA stimulated endothelial enhancers and requires p300 activity

**Contributing Authors:** Daniel S. Day, Bing Zhang, Joshua W. Ho, Lingyun Song, Jingjing Cao, Danos Christodoulou, Jonathan G. Seidman, Gregory E. Crawford, Peter J. Park and William T. Pu

**Contribution:** All analyses here are my own work, unless otherwise noted. Some figures for the analyses were generated by Dr. Bing Zhang but the underlying result was based on my analysis and work.

**Manuscript status:** Adapted from Zhang\*, Day\* *et al.*, *Genome Research*. 2013. 23: 917-927 (\* co-first author)

### 2.1 Abstract

Histone modifications are now well-established mediators of transcriptional programs that distinguish cell states. However, the kinetics of histone modification and their role in mediating rapid, signal-responsive gene expression changes has been little studied on a genome-wide scale. Vascular endothelial growth factor A (VEGFA), a major regulator of angiogenesis, triggers changes in transcriptional activity of human umbilical vein endothelial cells (HUVECs). Here, we used chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) to measure genome-wide changes in histone H3 acetylation at lysine 27 (H3K27ac), a marker of active enhancers, in unstimulated HUVECs and HUVECs stimulated with VEGFA for 1, 4, and 12 h. We show that sites with the greatest H3K27ac change upon stimulation were associated tightly with p300, a histone acetyltransferase. Using the variation of H3K27ac as a novel epigenetic signature, we identified transcriptional regulatory elements that are functionally linked to angiogenesis, participate in rapid VEGFA-stimulated changes in chromatin conformation, and mediate VEGFA-induced transcriptional responses. Dynamic H3K27ac deposition and associated changes in chromatin conformation required p300 activity instead of altered nucleosome occupancy or changes in DNase I hypersensitivity. p300 activity was also required for a

subset of dynamic H3K27ac sites to loop into proximity of promoters. Our study identified thousands of endothelial, VEGFA-responsive enhancers, demonstrating that an epigenetic signature based on the variation of a chromatin feature is a productive approach to define signal-responsive genomic elements. Further, our study implicates global epigenetic modifications in rapid, signal-responsive transcriptional regulation.

## **2.2 Introduction**

Genome-wide profiling of chromatin components between different cell types has demonstrated that transcriptional regulatory elements are decorated by characteristic patterns of post-translational histone modifications and other chromatin features and that these features contribute towards identifying cell type-specific gene regulation [49, 58, 82, 157], if possibly contributing towards the function itself. Such epigenetic signatures have been used to functionally annotate transcriptional regulatory elements that distinguish different cell types. For example, active enhancers, genomic elements that stimulate gene transcription, are marked by acetylation of histone H3 at lysine 27 (H3K27ac)[49, 67, 125, 158], the presence of the chromatin regulator p300[67, 87], hypersensitivity to nuclease digestion[72], and expression of RNA transcripts known as eRNAs[81]. However, much less is known about how chromatin signatures change during rapid cellular responses to extracellular cues and the effectiveness of epigenetic profiles in identifying transcriptional elements that mediate signal-responsive changes in gene expression.

Here, these questions are addressed by studying rapid, signal-responsive changes in chromatin features using vascular endothelial growth factor A (VEGFA) stimulated endothelial cells as a model system. Blood vessels nourish nearly every organ. Their growth is tightly regulated, and inadequate, excessive, or abnormal blood vessel growth is linked to a panoply of diseases, including ischemic heart disease, blinding eye diseases, and cancer[152]. A central regulator of blood vessel growth is vascular endothelial growth factor A. In response to VEGFA signaling, endothelial cells dramatically change their phenotype and gene expression profile[159]. Intracellular signaling downstream from VEGFA has been studied in depth, but relatively less is known about the transcriptional regulatory elements that respond to VEGFA signaling. We profiling of activating chromatin epigenetic marks were done in a 12-h time course of endothelial cell stimulation by VEGFA. We show that temporal variation of H3K27ac is a novel epigenetic signature that identifies VEGFA-regulated enhancers and predicts VEGFA responsive gene expression. This work further shows that the catalytic activity of p300 is required for dynamic changes in H3K27ac occupancy, altered chromatin architecture, and regulation of gene expression by VEGFA.

## **2.3 Results**

### **2.3.1 VEGFA Stimulation Induces Local H3K27ac Changes**

To study transcriptional and epigenetic regulation during angiogenesis, we treated human umbilical vein endothelial cells (HUVECs) with VEGFA and measured H3K27ac chro-

matin occupancy genome-wide at 0 (unstimulated), 1, 4, and 12 hours using ChIP-seq (see Figure 2-1). Our H3K27ac signal correlated strongly genome-wide across all time points, suggesting that there are few broad genome-wide changes in H3K27ac in response to VEGFA ( $r = .99$ , Pearson correlation; see Figure 2-2). Additionally, each H3K27ac data time point correlated with ENCODE H3K27ac ChIP-seq data for HUVEC ( $r = .45$ , Pearson correlation; see Figure 2-2). However, the correlation against ENCODE was not as strong as within the four time points, potentially attributable to differences in growth conditions since HUVECs are generally grown in medium with VEGF and other growth factors.

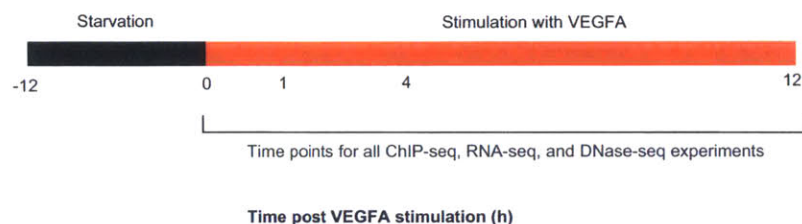


Figure 2-1: The design of the experiment. HUVECs were “starved” from their typical growth medium for 12 hours before being stimulated with VEGFA as a way to simulate angiogenesis. At the following non-negative time points, each of the described experiments below (*e.g.*, RNA-seq, ChIP-seq) was performed to capture the global landscape of the HUVECs at that point in the time course.

Closer inspection of the H3K27ac ChIP-seq data revealed focal regions with substantial changes in H3K27ac as a result of VEGFA stimulation (Figure 2-3). In order to identify regions with VEGFA-induced variation in H3K27ac, we estimated the H3K27ac mean-normalized variance (hereafter referred to as the “variance score”) across the 12 hour time course within a 200 bp sliding windows. This approach captured thousands of regions with substantial VEGFA-induced variation (see Figure 2-4). Out of sites with a log<sub>2</sub> variance score greater than 3, eight sites were selected by their location to eight genes implicated in angiogenesis for validation by ChIP-qPCR (performed by Dr. Bing Zhang). Sites near six genes were successfully assayed, and, in all six cases, the ChIP-qPCR results were consistent with the ChIP-seq data[150].

### 2.3.2 A dynamic VEGFA-regulated H3K27ac signature is tightly linked to p300 chromatin occupancy

The transcriptional coactivator p300 acetylates histones[162] and occupies tissue-specific enhancers[87]. To define the relationship of p300 chromatin occupancy to dynamic H3K27ac sites, p300 chromatin occupancy was measured during the VEGFA stimulation time course by ChIP-seq. The p300 ChIP-seq data were reproducible genome-wide between independent biological replicates[150] ( $r > 0.92$ , Pearson correlation). Comparison to publicly available ENCODE p300 occupancy data for the immortalized B-cell cell line GM12878[163] indicated that p300 binding was largely cell type-specific from its hematopoietic rela-

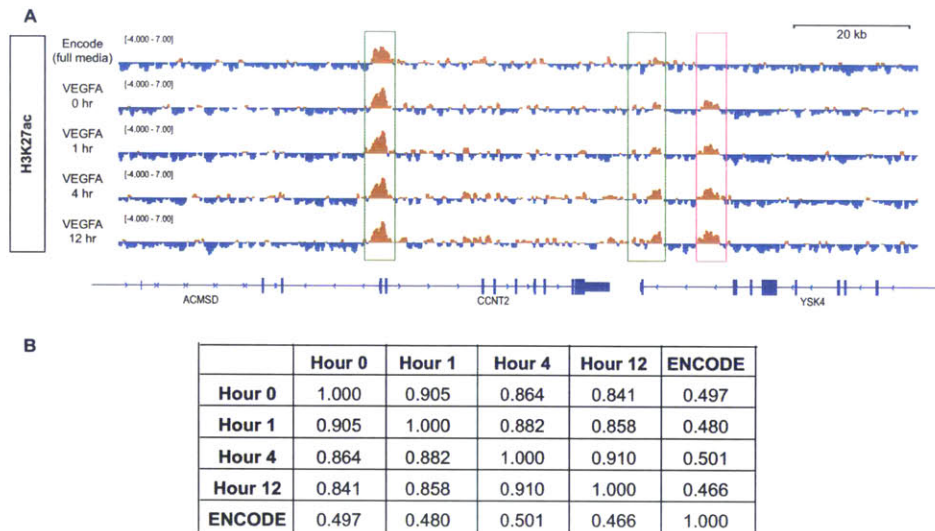


Figure 2-2: When comparing H3K27ac ChIP-seq profiles from this experiment to the ENCODE HUVEC data set, there was a strong correlation but notable differences between the experiments. In panel (A), this is a representative screenshot where many sites that are enriched in H3K27ac in this time course are also enriched in the ENCODE HUVEC data (green boxes). However, certain sites lack H3K27ac signal in the ENCODE data whereas there is signal across all 4 time points in the time course, suggesting that there are potential genetic (since the individuals from which the HUVECs were derived vary) and regulatory differences driven by changes in environment and growth conditions. In panel (B), there is a table of genome-wide Pearson correlations across each time point and against the ENCODE HUVEC H3K27ac signal. The strongest genome-wide correlation between the ENCODE data and the time points was at Hour 4. This makes sense since the ENCODE HUVECs are grown in VEGF-rich medium so they are, in a sense, always under signalling stress, and Hour 4 is the closest profile near that state relative to the other 3 time points.



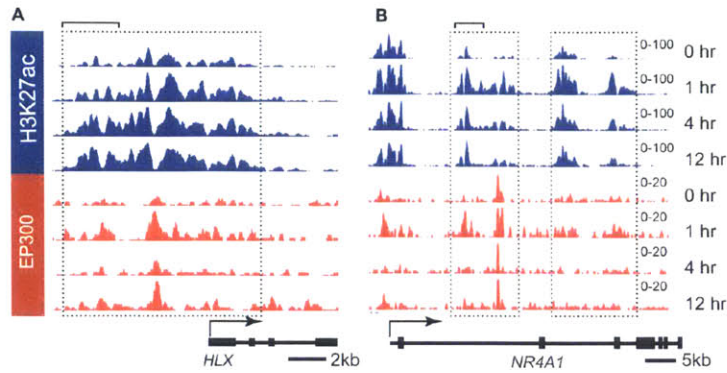


Figure 2-3: After VEGFA stimulation, H3K27ac signal (red track) changed locally over time. In (A) within the dashed box, H3K27ac signal near the *HLX* gene continually increased throughout the course of the experiment. In (B), we observed that multiple H3K27ac sites near the *NR4A1* rapidly increased after an hour but reduced in signal strength after 1 h. Hence, H3K27ac signal both responded to VEGFA but also had multiple temporal patterns. In both cases, variant H3K27ac sites were near p300 binding sites (blue track). Both *HLX* and *NR4A1* are genes that encode transcription factors that are upregulated during angiogenesis in response to VEGFA[160, 161], making sense that we observe significant H3K27ac change at these locations.

tive[150]. Next, from the VEGFA time course data, the distance between each H3K27ac window and nearest p300 binding site was calculated and stratified by the variance score of the H3K27ac window. Interestingly, most of the regions with the greatest H3K27ac variance scores occurred within 2 kb of p300 sites, while less variant H3K27ac sites tended to be further from p300 (see Figure 2-5). This result supported a tight relationship between p300 and dynamic but not static H3K27ac marks.

The enrichment of p300 near dynamic H3K27ac sites (see Figures 2-3 and 2-5) suggested that p300 is functionally involved in deposition of H3K27ac in response to VEGFA stimulation. To test the functional requirement for p300 in VEGFA-stimulated deposition of H3K27ac, the effect of p300 knockdown on H3K27ac chromatin occupancy was measured. siRNA p300 knockdown blocked VEGFA-stimulated deposition of H3K27ac at *NR4A1*, *HLX*, and *KDR*[150]. To determine if p300 histone acetyltransferase catalytic activity, as opposed to other functions mediated by p300 (*e.g.*, coactivator complex formation through protein-protein interactions), is required for VEGFA-stimulated H3K27ac, C646, a small molecular inhibitor of p300[164], pretreatment for 30 min of HUVECs blocked p300 enzymatic activity and VEGFA-stimulated deposition of H3K27ac[150].

Next, the extent to which p300 activity is required for dynamic H3K27ac deposition genome-wide was interrogated by performing H3K27ac ChIP-seq on cells pretreated with C646 and then stimulated with VEGFA for 0, 1, and 4 h. The cells became unhealthy by 12 h, precluding analysis at this timepoint. Focally, there was a massive reduction in H3K27ac variance when comparing the H3K27ac profiles (see Figure 2-6). When recalculating the variance scores for the control and C646 treated experiments using just the first three time points, p300 inhibition caused widespread reduction in H3K27ac variation induced by VEGFA (see Figure 2-7). However, some VEGFA-stimulated changes

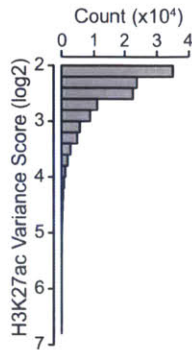


Figure 2-4: The distribution of number of sliding windows with a log<sub>2</sub> variance score of at least 2 (*i.e.*, at least four-fold variance to its mean).

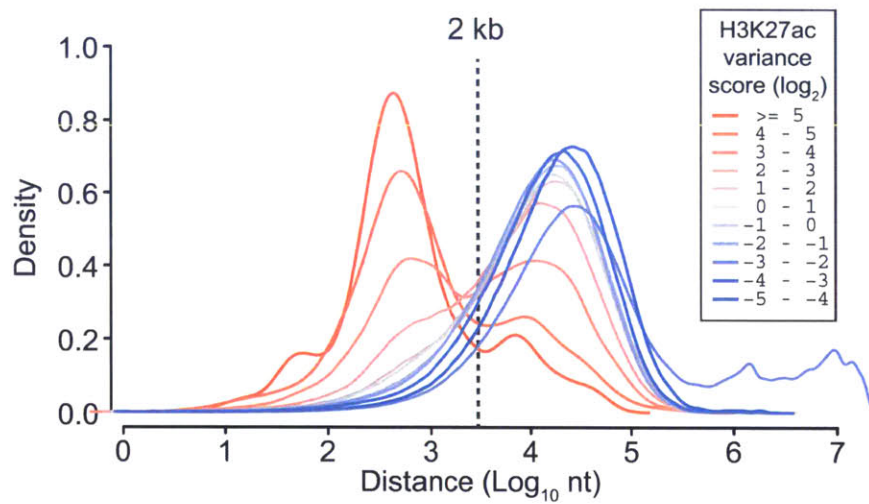


Figure 2-5: When measuring the distance between each H3K27ac sliding window and nearest p300 binding site, the sites with the higher variance score tended to be closer to p300 sites than H3K27ac regions with a lower variance score.



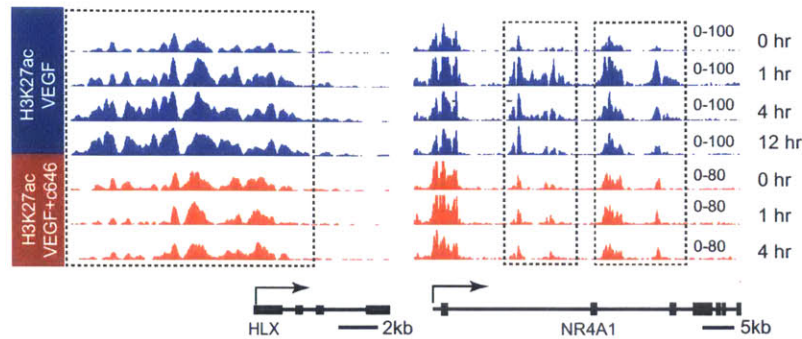


Figure 2-6: H3K27ac profile with and without C646 treatment to compare changes in H3K27ac upon blocking p300 acetyltransferase activity. In the regions highlighted previously (see Figure 2-3), there was a marked decrease in H3K27ac variability in the C646 treated cells. Additionally, it appears that the level of H3K27ac also goes down upon C646 as well.

in H3K27ac persisted, indicating that additional mechanisms also contribute to H3K27ac changes induced by VEGFA. Overall, these data indicate a key role of p300 in contributing to H3K27 acetylation induced by VEGFA.

Changes in nucleosome occupancy might have explained changes in H3K27ac, or changes in another histone modification may suggest H3K27ac was not the primary target of histone modification, but we generally ruled out both these possibilities. Changes in nucleosome positioning were previously reported to underlie rapid changes in the occupancy profile of H3K4me2[66]. We tested the hypothesis that changes in nucleosome occupancy contribute to the observed dynamic changes in H3K27ac by measuring total histone H3 and H3K4me2 occupancy at six dynamic H3K27ac sites[150]. We did not observe significant changes in histone H3 or H3K4me2 occupancy at any of these sites, indicating that acetylation of histone H3 rather than shifts in its position cause altered H3K27ac occupancy[150].

### 2.3.3 Temporal clustering of H3K27ac variation defined groups of chromatin regions with distinct function annotations and enriched transcription factor motifs

In order to investigate the significance of H3K27ac variation in response to VEGF stimulation, H3K27ac sites within 2 kb of a p300 peak in any time point were focused on, given p300's apparent role in acetylating these sites. This was also to improve the confidence in the H3K27ac sites studied given the relationship between H3K27ac and p300. In order to study the most variant sites, the H3K27ac window with highest variance score within 2 kb of a p300 peak were accumulated (removing any duplicate instances of the same window when the window may be near two different p300 peaks). In order to reduce the chance of false positives, the focus was on the top 20th percentile of windows ranked by their variance score (see Methods). Row scaling the H3K27ac enrichment followed by hierarchi-

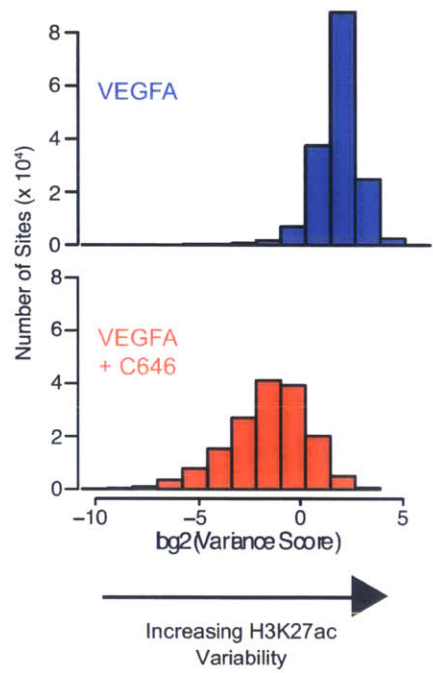


Figure 2-7: In C646 pre-treated cells that inhibits p300 acetyltransferase function, H3K27ac variability is dramatically decreased when comparing the distribution of H3K27ac variance scores.



cal clustering showed that H3K27ac enrichment at these sites followed three predominant temporal patterns (See Figure 2-8). We labeled these clusters as H1 (peak H3K27ac signal at 1 h; 4689 regions), H4-12 (peak H3K27ac signal at 4&A12 h; 3947 regions), and H0 (decreased H3K27ac signal at 4&A12 h; 3601 regions). Plotting H3K27ac signal intensity for each region illustrated the significant dynamic changes of H3K27ac binding in each temporal cluster (see Figure 2-9); Supplemental Fig. 7A). Cluster H4-12 was particularly interesting, because it showed initial depletion of H3K27ac signal at the peak center at 0 and 1 h and subsequent “filling-in” of the depleted region at 4 and 12 h (see Figure 2-9).

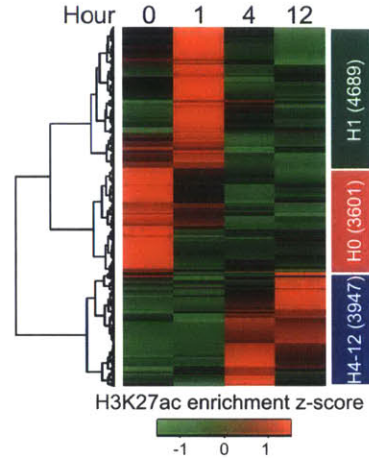


Figure 2-8: When analyzing the selected most variant H3K27ac windows (see Methods), row scaling the H3K27ac enrichment and performing hierarchical clustering revealed three dominant temporal patterns: where enrichment peaks at Hour 1 (H1), another at Hour 0 (H0), and the third at either Hours 4 and/or 12 (H-12). The number of H3K27ac variant windows that fell into each of these three patterns are indicated by their cluster.

The genome-wide analysis of the effect of the p300 inhibitor C646 on VEGFA-stimulated H3K27 acetylation showed a mechanistic requirement for p300 at the majority of sites. We, therefore, examined the C646 effect on the H1, H4-12, and H0 clusters in detail. Consistent with its essential role in VEGFA-stimulated deposition of H3K27ac, C646 strongly blunted H3K27ac accumulation in the H1 and H4-12 clusters (see Figure 2-9). Interestingly, the down-regulation of H3K27ac seen in the H0 cluster was also blunted by C646, suggesting secondary effects on counter-regulatory mechanisms that remove H3K27ac marks. When the regions were centered on p300 enrichment, aggregation plots of H3K27ac signal showed that maximal H3K27ac signal variation occurred adjacent to, rather than overlapping, p300[150]. Prior work showed that the chromatin landscape at most transcription factor binding sites is asymmetric. When we applied an algorithm for function strand segregation[165], we found that H3K27ac and p300 occupancy were both asymmetric in the H1, H4-12, and H0 clusters[150]. Interestingly, the distribution of H3K27ac and p300 with respect to the peak center was largely concordant, consistent with a mechanistic role of p300 in establishing the H3K27ac marks. p300 aggregation plots showed that p300 binding also changed during the VEGFA-stimulation time course[150]. We confirmed VEGFA-

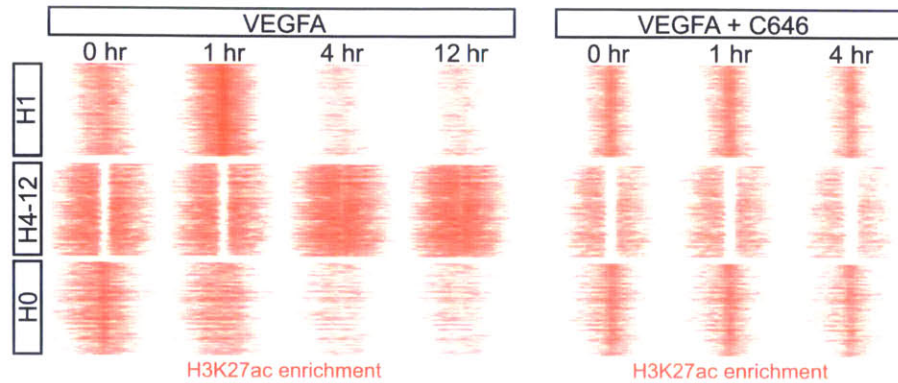


Figure 2-9: The most variant sites for H3K2ac were identified, row-scaled to emphasize time dependent patterns, and then grouped by hierarchical clustering.

stimulated enrichment of p300 by ChIP-qPCR[150]. For cluster H4-12, on average, p300 binding increased before H3K27ac occupancy. At cluster H1, these events appear to occur concurrently, suggesting that either the sequence of events differs at 1 h or that the data did not contain sufficient temporal resolution to order events peaking at this early time point. For cluster H0 with decreasing H3K27ac signal, p300 signal increased at 1 h, but H3K27ac signal did not[150], (Fig. 3B&A; Supplemental Fig. 7D), suggesting that other factors, such as increased HDAC activity, impeded H3K27ac deposition at these regions.

The location and function of the dynamic was further characterized, p300-associated H3K27ac sites. Most of these sites were located distal to transcriptional start sites (TSSs) of genes, consistent with the reported predominant location of p300[67, 87]. However, a significantly greater proportion of sites in cluster H1 were located in promoters, near gene TSSs ( $p < 10^{-46}$ , proportion test), while a significantly greater proportion of sites in cluster H4-12 were located in intergenic regions ( $p < 10^{-10}$ , proportion test)[150]. Dynamic, p300-associated H3K27ac sites were associated with 8454 adjacent genes. The majority of these genes did not overlap between temporal H3K27ac clusters[150]. Gene Ontology (GO) analysis showed that these different gene sets have distinct functional properties (see Figure 2-11). Both of the late-responding clusters, H4-12 and H0, were strongly enriched for terms related to vascular development, endothelial differentiation, and angiogenesis. In contrast, the early-responding H1 cluster was enriched for terms related to cell morphology, protein metabolism, response to oxygen levels, and TGFbeta receptor signaling, which are relevant to cellular stress responses in many cell types, including endothelial cells. Collectively, the data indicates that each temporal cluster of dynamic p300-associated H3K27ac sites was linked to regulation of varying aspects of cellular function. To identify transcription factors that regulate p300 recruitment and VEGFA-regulated H3K27 acetylation, we searched for overrepresented transcription factor binding motifs among sequences at p300 peaks in each cluster. *De novo* motif discovery revealed highly significant enrichment of ETS, FOX, AP1, STAT, and SP1 families in all three clusters. To further validate the motif discovery results, we performed ChIP-seq for ETS1 and found that 51% of p300-bound regions were co-occupied by ETS1 (see Figure 2-12).

The analysis also identified transcription factor motifs that occurred in one or two of the



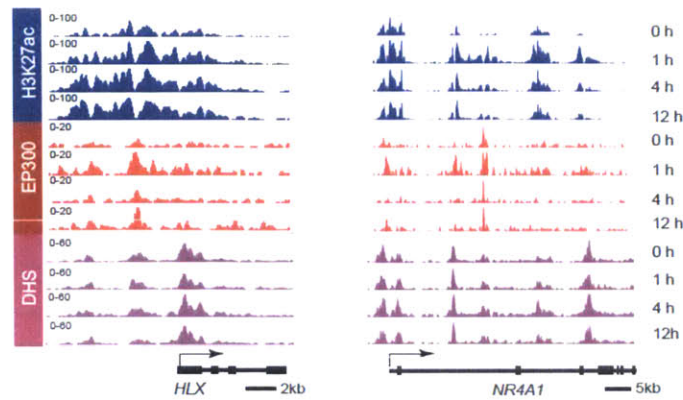


Figure 2-10: Analyzing regions near key angiogenesis regulations where we identified H3K27ac variation in response to VEGFA (see Figure 2-3), the DNase I hypersensitivity signal was plotted alongside these regions to analyze for changes in the openness of the chromatin environment around these variant sites. Here, there were no observed changes in DNase I hypersensitivity upon VEGFA stimulation. These sites had an open chromatin configuration prior to VEGFA stimulation based on positive DNase I hypersensitivity at Hour 0, suggesting these regions may have been “primed” in an open configuration at baseline ready to help facilitate VEGFA response.

H3K27ac clusters. The motif of ATF and CREB1, mediators of immediate early responses in multiple cell types (Altarejos and Montminy 2011), was significantly enriched in the H1 cluster. Also notable was overrepresentation of the SMAD binding motif in the H1 cluster, where we observed enrichment for the TGF- $\beta$  receptor signaling pathway (see Figure 2-13). We found over-representation of GATA and TEAD binding motifs (see Figure 2-13) in the H0 and H4-12 clusters, while the binding motif of RBP/J, the nuclear target of Notch signaling, was enriched in the H0 and H1 clusters (see Figure 2-13). These results suggest that members of these transcription factor families are important in orchestrating p300 recruitment and H3K27ac deposition in response to VEGFA stimulation. Enrichment of TF motifs at dynamic H3K27ac sites suggested that these TFs recruit p300 and thereby contribute to changes in H3K27ac. To test this hypothesis, we knocked down ETS1 or C-JUN (a component of the AP1 heterodimer) and measured the effect on dynamic H3K27ac sites directly bound by these factors. Validation experiments demonstrated efficient ETS1 or C-JUN knockdown in HUVECs after siRNA transfection[150] and corresponding reduction of ETS1 or JUN (also known as c-Jun) occupancy of tested dynamic H3K27ac sites (two sites tested per factor)[150]. This reduction of ETS1 or JUN binding attenuated H3K27ac changes at these sites in response to VEGFA[150]. These results suggest that TFs with enriched motifs are functionally important in mediating dynamic H3K27ac changes in response to VEGFA.

Cluster H1	
Ontology	Adj. P (-log10)
actin cytoskeleton organization	19
actin filament-based process	18
neg. reg. of protein metabolic process	16
TGFβ receptor signaling pathway	15
response to oxygen levels	15
cellular component disassembly	13
nuclear transport	12
I-kappaB kinase/NF-kappaB cascade	10

Cluster H4-12	
Ontology	Adj. P (-log10)
vasculature development	41
blood vessel development	40
angiogenesis	33
regulation of cell migration	23
regulation of endothelial cell proliferation	20
neg. reg. of MAPKKK cascade	13
endothelium development	12
endothelium differentiation	11

Cluster H0	
Ontology	Adj. P (-log10)
angiogenesis	15
regulation of endothelial cell migration	10
myeloid cell differentiation	9
I-kappaB kinase/NF-kappaB cascade	8
vasculogenesis	8
reg. of endothelial cell proliferation	8
reg. of smooth muscle cell proliferation	8
pos. reg. of Rho GTPase activity	6

Figure 2-11: Using GREAT[166], each set of H3K27ac variant sites per cluster were analyzed for functional Gene Ontology (GO) enrichments based on nearby genes. Representative GO terms are listed with corresponding p-values here.

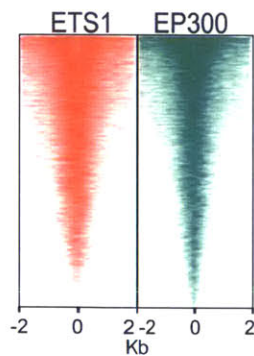


Figure 2-12: ETS1 co-occupies approximately 51% of regions bound by p300 as seen by this tag heatmap centered on all p300 binding sites over time with ETS1 enrichment also at the p300 binding site in the same position.

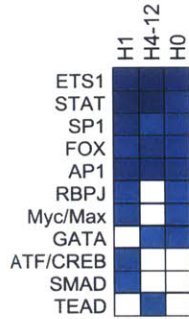


Figure 2-13: Enrichment of identified transcription factor motif families within each H3K27ac variant site.

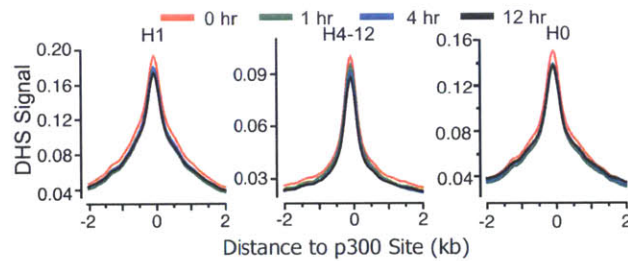


Figure 2-14: The DNase I hypersensitivity signal (DHS signal) under the p300 associated with its respective H3K27ac variant sites showed that all the enhancers had open chromatin at these sites at baseline and throughout the time course. However, there was no noticeable change on average with the signal levels over time suggesting these sites are “primed” for regulatory activity.

### 2.3.4 The dynamic H3K27ac signature defines VEGFA-responsive transcriptional regulatory elements

Active transcriptional regulatory elements are characterized by hypersensitivity to digestion by DNase I[99]. To further investigate whether dynamic, p300-associated H3K27ac sites are activating transcriptional regulatory elements, DNase I hypersensitivity followed by high-throughput sequencing (DNase-seq) was performed genome-wide during the VEGFA-stimulation time course [167]. Biological duplicate samples showed that the technique is highly reproducible[150]. Dynamic, p300-associated H3K27ac loci were DNase I hypersensitive (see Figure 2-14), consistent with their function as active transcriptional regulatory elements. Interestingly, on average, these regions did not change significantly in their sensitivity to DNase I digestion during the VEGFA-stimulation time course (see Figure 2-14), suggesting that most of these sites are already “open” and poised to respond to VEGFA stimulation. These data are consistent with our observation that H3K27 acetylation was not associated with changes in nucleosome occupancy[150].

Active enhancers are also characterized by production of transcripts known as eR-



NAs[81]. To further characterize the dynamic, p300-associated H3K27ac loci and confirm their enhancer activity, eRNA transcript levels were measured from the H3K27ac VEGFA responsive clusters (H1 and H4-12) by qRT-PCR (Fig. 5B; Supplemental Fig. 10A), testing 11 regions distal to gene bodies. In regions belonging to cluster H1, eRNA was strongly up-regulated at 1 h and then decreased at 4 and 12 h. In regions belonging to cluster H4-12, eRNA was up-regulated to maximal levels by 1 h and was sustained through hours 4 and 12. As controls, we measured eRNA transcripts from nearby regions with H3K27ac enrichment that did not change during the VEGFA time course. Although some control regions also showed VEGFA-stimulated changes in transcript level, their number and overall fold increase were less than at H3K27ac dynamic regions[150]. H4-12 cluster regions showed increased p300 recruitment and eRNA activity at 1 h, whereas H3K27ac did not increase until 4–12 h. To determine if p300 activity was required for VEGFA stimulated increases in eRNA, the acetyltransferase activity of p300 activity was blocked with C646 treatment, and the eRNA levels were remeasured. Acute p300 inhibition blocked VEGFA-stimulated up-regulation of eRNA[150]. In the H1 cluster, our experiments were unable to resolve temporal differences in p300 recruitment, eRNA activity, and H3K27ac binding, which concurrently peaked at 1 h. Nevertheless, p300 inhibition also blocked eRNA up-regulation at H1 regions, suggesting a similar role of p300 in this cluster. Our results suggest that p300 recruitment and acetylase activity are required for eRNA synthesis and precedes H3K27 acetylation.

Since the dynamic p300-associated H3K27ac loci had chromatin features of transcriptional regulatory regions, luciferase reporter assays were used to measure the transcriptional activity of 1- to 2-kb regions centered on 38 dynamic p300-associated H3K27ac loci. An equal number of loci were arbitrarily selected from each cluster, and tested regions were further subdivided into those located in promoter or nonpromoter regions. After reporter plasmid transfection, HUVECs were treated with VEGFA or vehicle. Luciferase activity measurements showed that dynamic, p300-associated H3K27ac regions belonging to H1 and H4-12 clusters activated transcription in response to VEGFA, while regions from the H0 cluster did not[150]. Regions belonging to the H1 cluster activated luciferase expression by 4 h, and expression then returned to baseline levels at 12 h. In contrast, regions from the H4-12 cluster increased luciferase activity at 4 h and maintained this through 12 h. Regions from promoter and nonpromoter regions behaved similarly[150]. These data indicate that regions in H1 and H4-12 clusters function as VEGFA-responsive transcriptional enhancers, while those in the H0 cluster with decreasing H3K27ac signal did not. Together, the DNase hypersensitivity, eRNA, and luciferase assays support VEGFA-responsive transcriptional enhancer activity of dynamic, p300-associated H3K27ac regions.

Next, temporal gene expression changes were connected with the dynamic, p300-associated H3K27ac loci. Gene expression profiling for transcript levels at 0, 1, 4, and 12 h after VEGFA stimulation was done by RNA-seq[150]. As expected, gene expression was highly dynamic following VEGFA stimulation, with 495 genes differentially expressed in at least one time point (cutoff by genes having  $q$  - value < 0.05)[150]. The RNA-seq data was validated by qRT-PCR and was generally concordant with previously reported microarray gene expression profiling data for HUVECs stimulated with VEGFA for 0 and 1 h[150, 159] and ENCODE HUVEC RNA-seq data[24, 150]. To evaluate the effect of the dynamic H3K27ac loci on VEGFA-regulated gene expression, we examined expression of

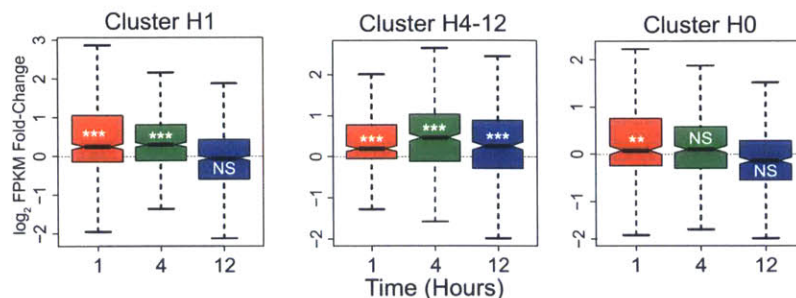


Figure 2-15: Based on differentially expressed genes in the VEGFA time course, changes in gene expression connected temporally with changes in nearby differentially expressed genes (assigning each H3K27ac variant site to the nearest differentially expressed genes within 100 kb). For example, in the H1 variant site cluster, where H3K27ac variability peaked at Hour 1, the nearby differentially expressed genes rapidly increased at Hour 1 falling through the rest of the time course. (\*\*  $p < .01$ , \*\*\*  $p < .001$ , NS = not significant, one sample Mann-Whitney U test)

genes that were differentially expressed and within 100 kb of dynamic, p300-associated H3K27ac sites (see Figure 2-15). For genes associated with the H1 cluster, transcript levels were significantly higher at 1 and 4 h compared to 0 h, and returned to baseline by 12 h. For genes associated with the H4-12 cluster, expression increased by 1 h, became further increased by 4 h, and was sustained through 12 h. Expression of H0-associated genes was slightly but significantly increased at 1 h but returned to baseline levels at 4 and 12 h. Thus, each cluster of H3K27ac variation was associated with a corresponding temporal pattern of altered gene expression. These data further support the activity of the dynamic H3K27ac loci in the H1 and H4-12 clusters as transcriptional enhancers.

### 2.3.5 Dynamic H3K27ac sites and p300 participate in VEGFA-stimulated chromatin looping

Enhancers are thought to stimulate transcription at promoters by forming chromatin loops[14]. Whether VEGFA rapidly stimulated chromatin looping at dynamic, p300-associated H3K27ac loci was investigated. The Mediator complex has been implicated in the formation of chromatin loops[14, 120]. MED1 and MED12, encoding Mediator complex subunits, were highly expressed in HUVECs[150]. MED1 and MED12 occupancy of dynamic, p300-associated H3K27ac sites was measured by ChIP-qPCR[150]. At seven of eight loci belonging to cluster H1, MED1 and MED12 enrichment strongly increased at 1 h of VEGFA treatment, then declined to basal levels at 4 to 12 h. In cluster H4-12, MED1 and MED12 were enriched at most loci, but the degree of enrichment did not change in a consistent temporal pattern following VEGFA treatment. These results indicate that the Mediator complex is bound to dynamic, p300-associated H3K27ac sites and suggest that these sites may undergo VEGFA-stimulated looping. To directly test the hypothesis that dynamic, p300-associated H3K27ac sites loop into proximity with promoters after VEGFA stimu-

lation, chromatin conformation capture[168] was used to study temporal changes in chromatin conformation involving three loci with VEGFA-stimulated increases in H3K27ac. Upstream of *DUSP5*, a dynamic H3K27ac site belonging to cluster H1 became transiently associated with the promoter at 1 h (see Figure 2-16), when it was maximally occupied by H3K27ac, p300, and MED1/12, and maximally transcribed as eRNA. At later time points, the association of these regions declined, coincident with decreased p300 and MED1/12 occupancy, and decreased eRNA transcription. Upstream of *KDR* (encoding VEGFR2, a VEGFA receptor), dynamic H3K27ac sites belonging to cluster H4-12 became associated with the promoter within 1 h of VEGFA stimulation (see Figure 2-16). This correlated with its time course of p300 and MED1/12 occupancy and eRNA transcription but preceded its maximal occupancy by H3K27ac. Similar observations were made at a second dynamic H3K27ac site from cluster H4-12 located upstream of the endothelial gene *CD34* (see Figure 2-16). Thus, at these sites, VEGFA stimulation rapidly altered chromatin conformation and stimulated eRNA transcription, and these events preceded deposition of H3K27ac.

To probe the requirement of p300 in chromatin looping, chromatin conformation capture experiments chromatin conformation capture experiments was repeated in the presence of the p300 inhibitor C646 (see Figure 2-16). C646 blocked VEGFA-stimulated chromatin looping, thereby establishing the importance of p300 in establishing chromatin loops. Consistent with a key role of p300 acetyltransferase activity in mediating VEGFA-stimulated chromatin changes and activation of gene transcription, C646 potentially blocked up-regulation of genes normally induced by VEGFA, including *DUSP5*, *KDR*, *NR4A1*, and *CD34*(see Figure 2-17).

## 2.4 Discussion

Epigenetic signatures define transcriptional regulatory elements that underlie the distinct gene expression programs of different cell types, and these signatures have been used to annotate cell type-specific functional elements[49, 50, 58, 82]. However, less is known about how the chromatin landscape responds to transient environmental cues. To gain insights into this area, we studied changes in H3K27 acetylation that occur within 12 h of endothelial cell stimulation with VEGFA, a major regulator of angiogenesis.

VEGFA induced rapid changes in H3K27ac at thousands of genomic loci. Some of the VEGFA-regulated transcriptional regulatory elements were defined by dynamic, temporal changes in H3K27ac. These regions had characteristics of activity-regulated enhancers: they were tightly linked to p300 chromatin occupancy, had functional annotations linked to blood vessel development, were transcribed as VEGFA-stimulated eRNAs, and engaged in VEGFA-regulated chromatin looping and gene expression. These regions with dynamic H3K27ac exhibited VEGFA-stimulated transcriptional activity in both luciferase assays and in HUVEC gene expression profiles, and p300 inhibition blocked VEGFA-induced changes in H3K27ac and gene expression. Thus, our study indicates that the epigenome is an integral participant in signal-induced transcriptional responses. We developed a novel epigenetic signature based on the signal-induced variation of H3K27ac chromatin occupancy. Using this signature, we identified thousands of novel endothelial, VEGFA-responsive transcriptional regulatory elements and the transcription factor families that are



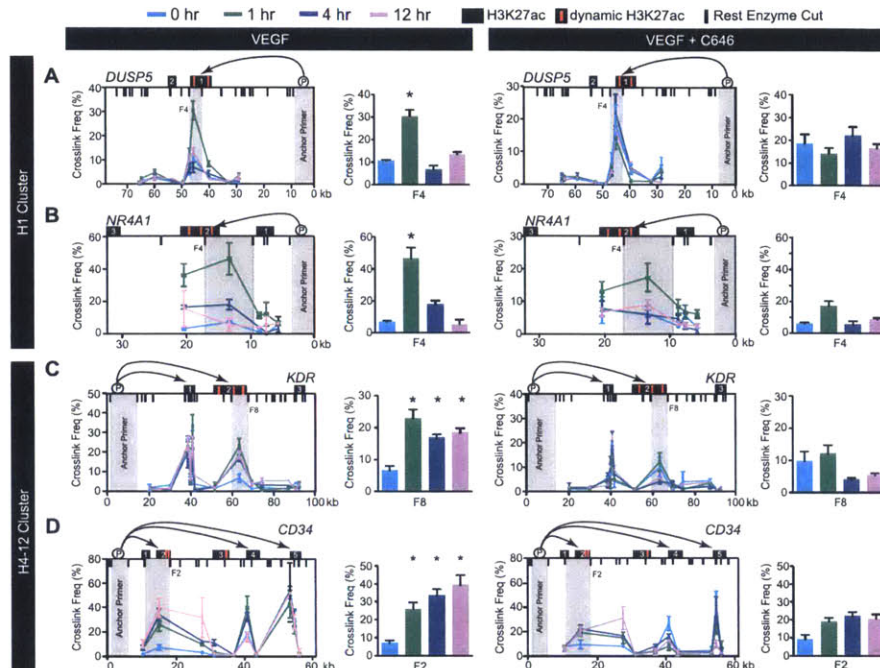


Figure 2-16: For tested variant H3K27ac sites, there was temporal changes in chromatin looping towards the target promoter in response to VEGFA. In each of the 4 sites tested (2 from the H1 cluster and 2 from the H4-12 cluster), there was specific changes in chromatin looping at our variant H3K27ac sites relative to other tested non-variant H3K27ac. p300's acetyltransferase activity was required for observed temporal changes because changes in chromatin looping were lost upon C646 treatment, which blocks p300 acetyltransferase. Also, notice that in each case there was initial looping of the variant site to the target promoter at Hour 0, suggesting at some of the chromatin is already ready for looping. (Result produced and provided by Dr. Bing Zhang.)

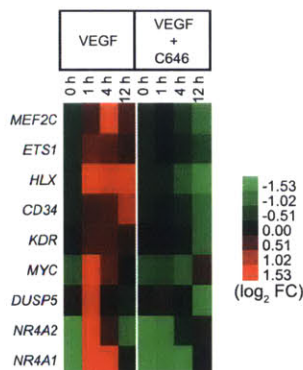


Figure 2-17: After blocking p300-mediated chromatin looping with C646, VEGF responsive gene expression change is ablated at selected genes (tested by qRT-PCR, provided by Dr. Bing Zhang).

likely to regulate them. These regulatory elements were associated with genes that regulate angiogenesis and could be separated into distinct functional groups based upon their temporal variation of H3K27 acetylation. These data will provide an important resource for future studies of the transcriptional regulation of angiogenesis, although we note that this study was performed in cultured venous endothelial cells and that other VEGFA-responsive endothelial enhancers active in other endothelial cell types or active *in vivo* were likely not detected in this system. More broadly, the potential application the epigenetic signature based upon signal-induced chromatin feature variation to other biological systems will enhance annotation of activity-regulated functional elements genome-wide. Previous studies suggested that nucleosome dynamics define activity-regulated transcriptional enhancers[66, 82]. However, the data suggest that rapid changes in H3K27ac were not due to changes in chromatin accessibility/ nucleosome occupancy. Rather, dynamic H3K27 acetylation was closely associated with p300, and, indeed, p300 and its acetyltransferase activity were required to write these marks. These data indicate that epigenetic enzymatic activity is also an important factor that establishes activity-regulated transcriptional enhancers. Our experiments highlight the crucial role of p300 in mediating signal responsive changes in H3K27ac and gene expression. p300 is a histone acetyltransferase that was previously reported to occupy tissue-specific transcriptional enhancers[87]. Our data show the proximity of regions occupied by p300 and regions with VEGFA-stimulated variation in H3K27ac. Inhibition of VEGFA-induced H3K27ac accumulation by p300 antagonists supports the causal role of p300 in dynamic variation of H3K27ac occupancy. Furthermore, p300 inhibition dramatically blocked gene expression changes induced by VEGFA. These data directly demonstrate the key role of p300 in executing VEGFA-induced transcriptional responses and suggest more broadly that p300 is required for signal-induced changes in histone acetylation and gene expression. To identify transcription factors that participate in the VEGFA transcriptional response and recruit p300 to dynamic H3K27ac sites, we found transcription factor motifs enriched in p300-bound regions. ETS, FOX, AP1, and STAT transcription factor motifs were enriched in all three clusters, suggesting that members of these transcription factor families broadly participate in VEGFA-driven transcriptional changes. The key role of several ETS factors in angiogenesis was reviewed recently[169]). We directly confirmed ETS1 occupancy of most p300- bound regions, validating the motif analysis and providing a resource for further study of the role of ETS1 in angiogenesis and VEGFA-induced gene expression changes. The extensive overlap between p300 and ETS1 binding suggests that ETS1 may contribute to p300 recruitment. Consistent with this hypothesis, ETS1 knockdown blocked VEGFA-induced H3K27ac changes at ETS1- bound loci. Compared to ETS, relatively less is known about the role of FOX, AP1, and STAT transcription factor family members as effectors of VEGFA signaling. Our data identify regions potentially regulated by these factors downstream from VEGFA. Recently, FOX transcription factors were reported to interact with ETS factors to regulate vasculogenesis, and similar interactions may also contribute to angiogenesis (De Val et al. 2008). Our data also indicate that AP1 is an important transcriptional effector of VEGFA. Although this role of AP1 has not been studied, AP1 is well-positioned in intracellular signaling pathways to act in this capacity: AP1 is a major nuclear target of MAPK signaling, which is robustly activated downstream from VEGFA. We also identified transcription factor motifs that were enriched in a subset of dynamic H3K27ac clusters, suggesting a link to specific temporal

patterns of H3K27 acetylation and to specific functional pathways. In the early-responding H1 cluster, we detected significant enrichment for the ATF1/CREB1 (activating transcription factor 1 and cyclic-AMP response element binding protein 1) motifs (see Figure 2-13). These transcription factors mediate immediate early responses, which predominate the functional terms linked to the H1 cluster. GATA and TEAD motifs were overrepresented in the H4-12 and H0 clusters. GATA2 has been implicated as a key regulator of endothelial gene transcription[170], and Tead4 (also known as RTEF-1 and TEF-3) was recently reported to be required for VEGFA-stimulated angiogenesis. GATA2 and Tead4 likely contribute to endothelial cell-specific functional term enrichment in the H4-12 and H0 clusters. GATA factors are also crucial in regulating hematopoiesis. However, GO terms related to blood development were not overrepresented in the H4-12 or H0 cluster, suggesting that the GATA motifs identified by our analysis are selectively active in endothelial cells. The H0 and H1 clusters, which share a decline of H3K27ac at 4h, were both enriched for the binding motif of RBP/J, the nuclear target of Notch signaling. Interestingly, VEGFA signaling activates the Notch pathway, which, in turn, antagonizes VEGFA action in an auto-regulatory loop[171, 172]. Collectively, these data suggest that distinct transcription factor families contribute to the different temporal and functional properties of the H0, H1, and H4-12 clusters.

In addition to its role in depositing H3K27ac, p300 catalytic activity was required for VEGFA-induced chromatin looping. Previous to this work, the requirement of p300 acetyltransferase function in chromatin looping does not appear to have been reported previously. Furthermore, the time course data, surprisingly, suggest that, for members of the H4-12 cluster, eRNA expression and chromatin looping occur prior to up-regulation of H3K27ac, yet are dependent on p300 catalytic activity. Further work is required to establish the mechanism(s) through which p300 acetyltransferase activity promotes eRNA expression and chromatin looping and how the interplay between these chromatin properties regulates gene transcription.

## 2.5 Methods

### 2.5.1 Experimental Work and Growth Conditions

Detailed descriptions of growth conditions and experimental work provided by Dr. Bing Zhang are listed in detail in Zhang *et al.*[150].

### 2.5.2 ChIP-seq Analysis, Peak Calling for p300 and DNase I Hypersensitivity Sites

ChIP-seq and DNase-seq reads were mapped with Bowtie (Langmead et al. 2009) (summarized in Supplemental Table 5). Peak calls for p300 ChIP-seq data were generated by using *spp*[63] with the *find.binding.positions* function using a 1% FDR cutoff. Peaks from DNase I hypersensitivity data were called using *f-seq*[72, 167].

### 2.5.3 Calculating the H3K27ac variance score

H3K27ac ChIP-seq and input reads for each time point separately were binned into 200 bp sliding windows at 50 bp increments. Within each bin per time point, the H3K27ac enrichment was calculated as follows:

$$E_b = \frac{c_b}{i_b} * \frac{I}{C}$$

Where  $E_b$  is the enrichment  $E$  in bin  $b$ ,  $c_b$  and  $i_b$  are the count of ChIP and input reads, respectively, in bin  $b$ , and  $C$  and  $I$  are the total number of mapped ChIP and input reads, respectively, for that time point. Bins with fewer than 10 total reads (*i.e.*,  $c_b + i_b < 10$  for bin  $b$ ) were considered to be too low of signal and removed from further analyses.

To calculate how variable H3K27ac was in bin  $b$  over the 4 time points, each bin that was not removed for low signal had its variance score calculated as follows:

$$VarScore_b = \frac{\sigma_b^2}{\mu_b}$$

Where  $\sigma_b^2$  is the variance of H3K27ac enrichment in bin  $b$  over time and, accordingly,  $\mu_b$  is the mean enrichment in bin  $b$  over time. We found bins with a low mean enrichment over time to be more noisy, so we only further analyzed bins with a mean enrichment over time of at least 3 (except in the one exception described below). For display purposes only, we often displayed the variance score of a bin on the  $\log_2$  scale, since it was a monotonic transformation and generally made the variance score values more manageable to read.

For each p300 site, we identified the H3K27ac region within 2 kb with greatest variance score. Of these regions, the top 20th percentile was defined as dynamic, p300-associated H3K27ac regions.

### 2.5.4 RNA-seq Analysis

RNA-seq and differential expression analysis was performed using tophat to align the RNA-seq reads to the Ensembl hg19 annotation followed by cufflinks and cuffdiff for FPKM quantification and calling of differentially expressed genes[173]). Cuffdiff was run with the “-T” option to run the analysis in time series mode such that the differential comparison would only be between successive samples passed to the program and not across all samples. Accordingly, the aligned RNA-seq BAM files were passed to cuffdiff in the order of the time course. Differentially expressed genes were called by using a 5% FDR cutoff.

### 2.5.5 Transcription factor motif analysis

Motif discovery was performed using DREME[174], and motifs were annotated with TOM-TOM[175].

## **2.5.6 Browser Views**

All browser views of the ChIP-seq and other high-throughput sequencing data were generated using Integrated Genomics Viewer (IGV) [176].

## **2.5.7 Data access**

All sequencing data were deposited in the NCBI Gene Expression Omnibus (GEO) (<http://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE41166. The data are also available through the Cardiovascular Development Consortium at <https://b2b.hci.utah.edu/gnomex/>. [150].

## Chapter 3

# Analyzing the Transcriptional Regulation of Endothelial Cells and VEGFA Stimulus-Response

**Contributing Authors:** Daniel S. Day, Bing Zhang, Peter J. Park, William T. Pu

**Contributions:** All analysis of the data presented within this chapter were performed by me.

**Manuscript Status:** *in preparation*

### 3.1 Abstract

Angiogenesis is a critical in both normal physiology and disease, but the transcriptional regulation behind this process is poorly understood. With advances in high-throughput sequencing technology, new methods enable the analysis of transcription factor binding genome-wide using chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq). By analyzing the dynamic change in co-binding of candidate VEGFA-responsive transcription factors identified from VEGFA-responsive regulatory elements, candidate enhancers through identified by co-binding of transcription factors were identified to show the complexities of VEGFA-induced transcription factor binding change and their relationship with VEGFA-responsive genes. Additionally, analysis of gene and chromatin structure revealed multiple novel insights into the transcriptional and chromatin related patterns driving angiogenesis.

### 3.2 Introduction

Angiogenesis involves multiple waves of transcriptional change in endothelial cells in order to form a new blood vessel[2, 150, 152, 159], but the regulatory network driving these changes is poorly understood[5]. Several transcription factors have been implicated in regulating angiogenesis[5, 177], and it is thought that angiogenesis-related regulatory elements require the co-binding of a variety of transcription factors to drive the process[5]. The

development of chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) enables the genome-wide assaying of transcription factors and other chromatin bound proteins. Previously, dynamic changes H3K27ac in combination with p300 was analyzed in vascular endothelial growth factor A (VEGFA) stimulated human umbilical vascular endothelial cells (HUVECs) to identify angiogenesis-related regulatory elements[150]. This analysis identified a number of candidate transcription factors that may regulate angiogenesis [150]. In order to understand the binding changes of these candidate transcription factors further, their dynamic binding pattern was analyzed during angiogenesis in combination with changes in VEGFA-induced gene expression and chromatin structure.

In this study, multiple transcription factors and chromatin marks were assayed using ChIP-seq over the HUVEC time course described previously[150]. This identified multiple VEGFA-induced transcription factor co-binding modules, including a novel GATA2-MYC co-binding cluster, and changes in the chromatin landscape at HUVEC enhancers. In addition to these analyses, the re-analysis of VEGFA-responsive gene expression data identified a large set of non-coding RNA genes differentially expressed in concert with differentially expressed protein-coding genes. Together, these analyses provide deeper insight into the regulation of angiogenesis than ever before.

## **3.3 Results**

### **3.3.1 Experimental Design**

In order to further study the transcriptional regulation of angiogenesis, several histone modifications and transcription factors were assayed using the same experimental setup as previously described[150] (see Figure 2-1). In addition to the previously assayed H3K27ac[150], the additional histone modifications assayed were chosen to further annotate the euchromatic and heterochromatic regions of the HUVEC genome (see Table 3.1). In order to assay the transcriptional network driving VEGFA-responsive gene expression change, 10 transcription factors or co-activators were assayed via ChIP-seq (see Table 3.1), including p300 and ETS1 from the previous study[150], to analyze for dynamic changes in transcription factor binding at candidate angiogenesis-related regulatory elements. The choice of transcription factors assayed was based on enriched transcription factor motifs within the previously identified, p300-based VEGFA-responsive regulatory elements[150] (see Figure 2-13) and whether the assayed transcription factor had been previously implicated in regulating angiogenesis[5].

### **3.3.2 Analysis of Temporal Pattern of VEGFA-responsive Protein Coding and Non-coding RNA Gene Expression Change**

Genes responsive to VEGFA, or VEGFA responsive genes, were defined by identifying genes with a significant fold-change in gene expression relative to Hour 0 (the baseline) using RNA-seq data previously generated[150]. The RNA-seq time course data was mapped against the human GENCODE transcript reference annotation[178]. A gene was VEGFA-responsive if it had a minimum absolute two-fold gene expression change relative to Hour

Chromatin Marks	Transcription Factors/Co-activators
H3K27ac	p300
H3K27me3	ETS1
H3K4me1	GATA2
H3K4me2	MYC
H3K4me3	JUN
H3K9me3	ERG
H3K36me3	FLI1
	NICD
	RBPJ
	Pol II

Table 3.1: The ChIP-seq generated samples of histone modifications, transcription factors, and co-activators analyzed in this study. All of these ChIP-seq data sets were generated by Dr. Bing Zhang.

0 in at least one time point and the following conditions also held: a gene A) was expressed (FPKM>1) in at least one time point, B) never had a FPKM of 0 at any time point, C) had at least one Pol II ChIP-seq peak call overlapping the gene body in at least one time point and D) had a minimum of 10 RNA-seq reads mapped to the gene at each time point. Genes passing these criterion were clustered via k-means clustering to identify temporal patterns of VEGFA-induced gene expression change. Using three clusters, the predominant temporal patterns that emerged were early upregulate genes (ERGs), late upregulated genes (LRGs), and downregulated genes (DRGs) (see Figure 3-1). The timing of gene expression changes in response to VEGFA was consistent with timing of signal-responsive genes in other cell types[132, 179]. Curiously, Each VEGFA-responsive gene set had a similar number of protein-coding and non-coding RNA genes (see Figure 3-2). A fourth set of non-VEGFA responsive genes (NRGs) was also defined for expressed (FPKM>1) genes but did not significantly change throughout the time course. Across the four gene sets, NRGs had both the largest number of and a higher proportion of protein-coding genes to non-coding RNA genes (see Figure 3-2).

The expression level and fold-change magnitude of each gene set was different from the other. When comparing absolute expression levels between sets, it suggested that all the VEGFA-responsive gene sets had lower expression on average than NRGs (see Figure 3-3), which may partially be attributed to the definition of a VEGFA-responsive gene. However, when comparing the magnitude of gene expression fold change, ERGs had the strongest absolute expression fold-change increase (see Figure 3-4). ERGs also tended to rapidly increase in expression and then fall back to baseline (see Figure 3-4). Comparatively, both LRGs and DRGs had a strong absolute magnitude in gene expression fold change, but they gradually changed in gene expression over time (see Figure 3-4). Hence, the timing of a gene's expression change in response to VEGFA determined how that gene was regulated over the time course.

These VEGFA-responsive gene sets were further analyzed for significant biological



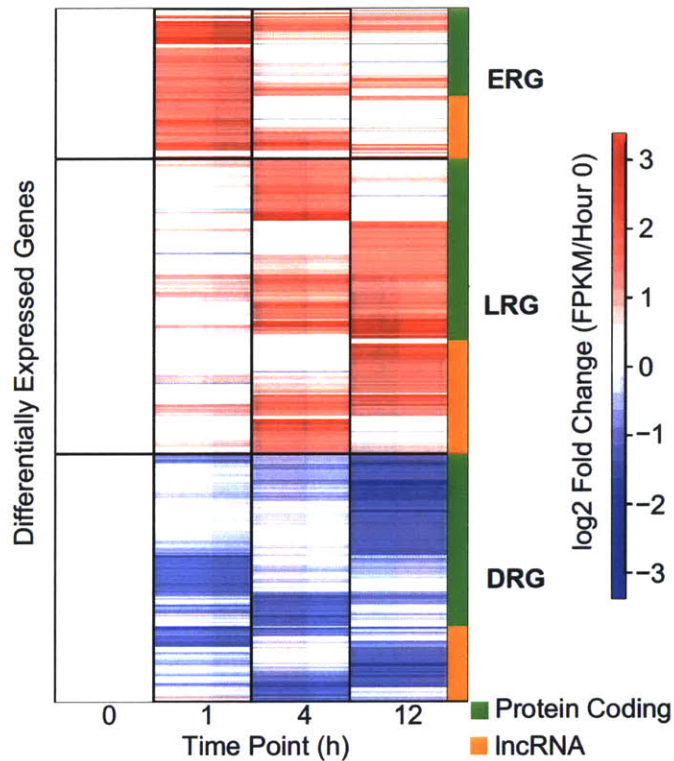


Figure 3-1: VEGFA-responsive genes, gene with a minimum two-fold gene expression change relative to Hour 0, were clustered to identify major temporal angiogenesis-related gene expression changes. The three main clusters were early responsive genes (ERGs), late responsive genes (LRGs), and downregulated genes (DRGs). Each clusters had several protein coding and non-coding RNA genes co-expressed in time in response to VEGFA.

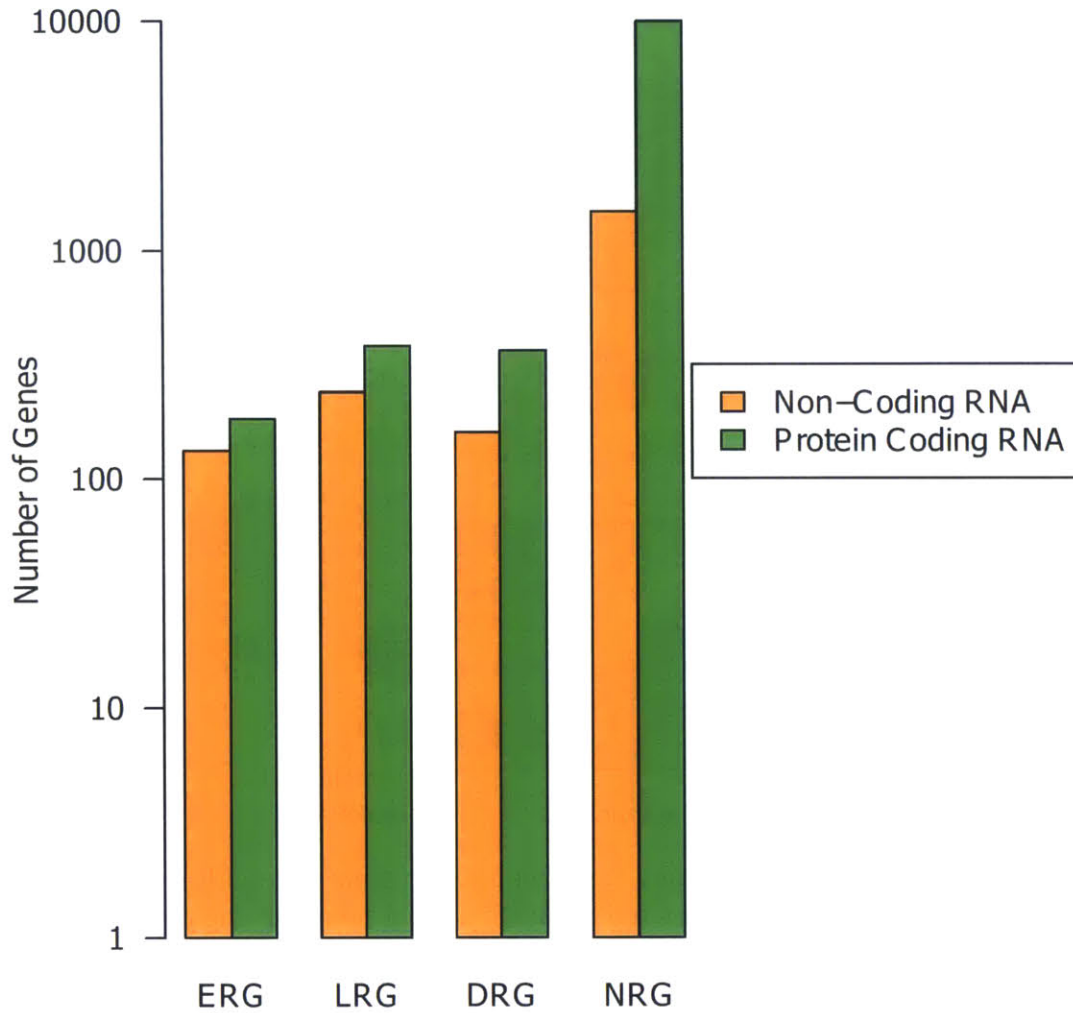


Figure 3-2: Protein coding and non-coding RNA gene count of each gene set. VEGFA-responsive gene sets (ERGs, LRGs and DRGs) had a balance of protein coding and non-coding RNA genes, but NRGs were predominately protein coding genes.

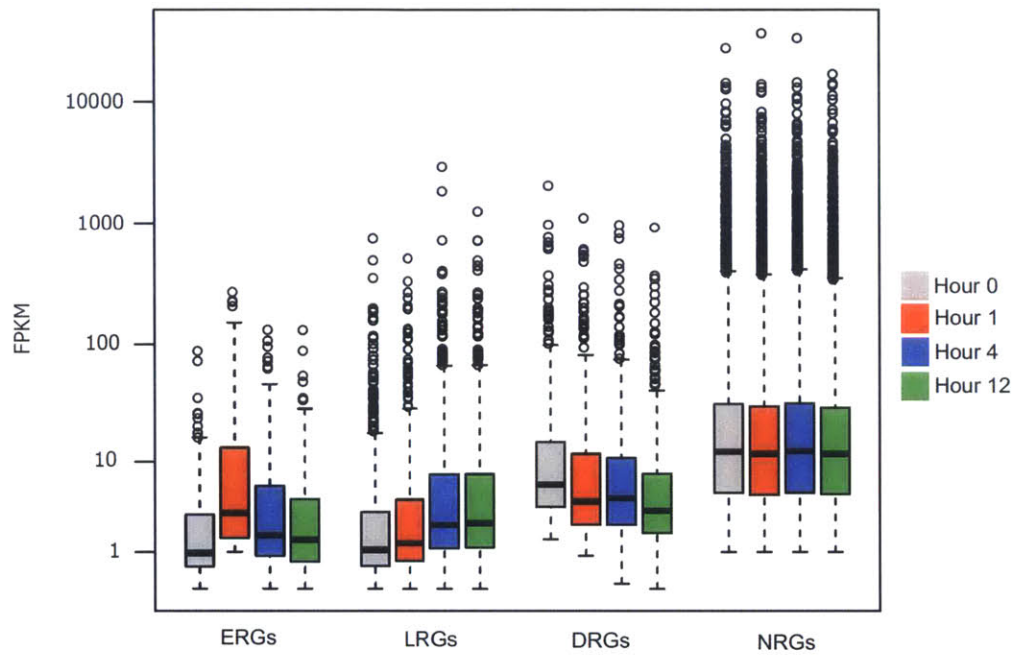


Figure 3-3: Boxplot of expression ranges (in FPKM) for each of the four gene sets per time point. All the VEGFA-responsive genes tended to have a lower median expression level than NRGs. ERGs tended to be only upregulated at Hour 1 before returning to baseline. LRGs gradually increased their expression level during the time course. Conversely, DRGs were gradually downregulated over the time course.

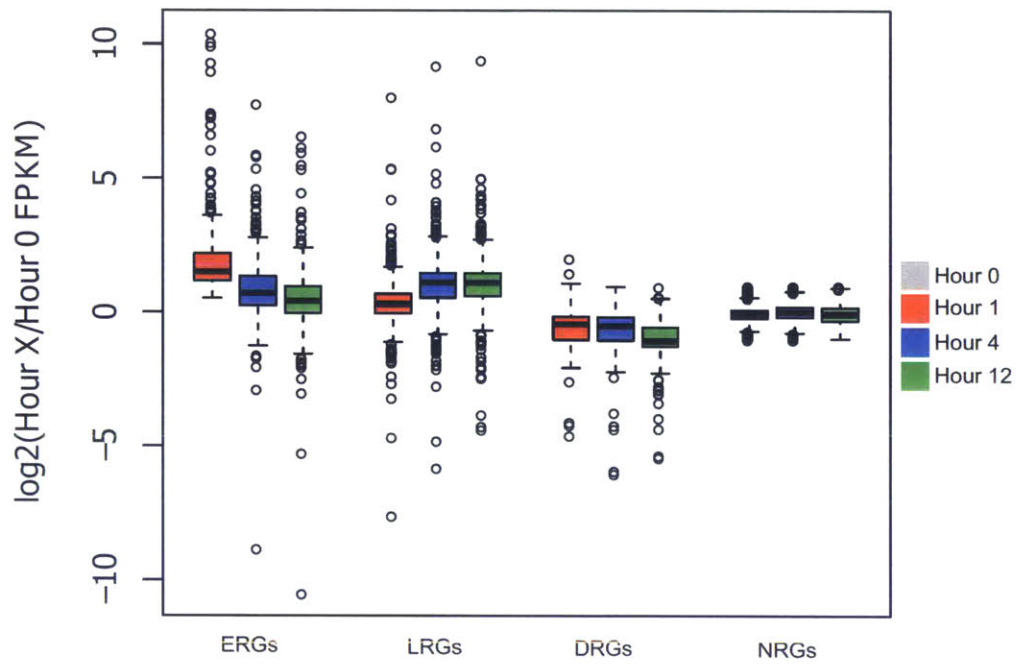


Figure 3-4: Expression fold change relative to Hour 0 over time at VEGFA-responsive and non-responsive expressed genes. ERGs showed the highest average fold change increase at Hour 1 with a rapid decline, similar to their change in average FPKM levels, compared to the gradual increase in expression by LRGs.

functions in gene set in order to both validate the analysis and identify novel biological functions of VEGFA-responsive genes. In order to do this, Gene Ontology (GO) enrichment analysis was performed on the protein-coding genes of each VEGFA-responsive gene set. Enriched GO biological process (BP) terms were identified (using a 10% FDR cutoff) against all expressed genes in HUVECs (see Methods). Enriched GO BP terms were clustered across the three responsive gene sets to identify functional overlaps (see Figure 3-5), and this also showed that enriched functions were consistent with previous studies. The ERGs had many enriched GO BP terms, which included functions involving cell proliferation and blood vessel development (see Figure 3-5). LRGs tended to be more uniquely enriched for blood vessel morphogenesis and cell adhesion. DRGs tended to be enriched in functions involving the cell cycle. Curiously, ERGs had the largest number of enriched GO BP terms (see Figure 3-5). This may be because ERGs have been more extensively studied and annotated compared to LRGs and DRGs, increasing their likelihood of having enriched GO BP terms.

The four gene sets had different gene structures from each other. When comparing the total gene size, this showed that ERGs were significantly smaller on average than the other three gene sets (see Figure 3-6), where the other gene sets generally had a similar gene size except for DRGs and NRGs (see Figure 3-6). These differences in the DNA gene size somewhat translated into differences at the RNA transcript level (see Figure 3-7). However, the average RNA transcript size of each gene set was more similar to each other than the differences at the gene level. This might suggest that ERGs were selected for a smaller size, possibly to facilitate the observed rapid upregulation in response to VEGFA.

Between each gene set there were significant differences in the average number of exons per gene (see Figure 3-8), using all unique annotated exons for each gene. In addition to their smaller gene size (see Figure 3-6), ERGs had fewer number of exons per gene on average than any of the other gene sets (see Figure 3-8). Surprisingly, NRGs had the highest average number of exons per gene (see Figure 3-8) even while it was similar in gene size to LRGs and DRGs (see Figure 3-6). Comparatively, LRGs and DRGs had a similar average number of exons per gene. Since this analysis was independent of the underlying expressed isoform(s), this suggests a distinct structural difference between genes in these sets. ERGs may be both shorter and have fewer exons to facilitate rapid Pol II transcription, but LRGs, DRGs, and NRGs may utilize alternative splicing more frequently.

### **3.3.3 Low Temporal Variability of Typical Chromatin Marks of Active TSSes at VEGFA Responsive Genes**

Previous studies have suggested that signal-responsive genes may undergo chromatin remodeling at their TSS in order to facilitate signal-responsive changes in Pol II initiation[132], but previous studies analyzed a small number of genes at a time. VEGFA-responsive genes may require chromatin remodeling at TSS in response to VEGFA as well. This may occur with chromatin marks known to occur at expressed TSSs[8]: H3K27ac, H3K4me3 and DNase I hypersensitivity. Over the time course, changes in these chromatin marks (including DNase I hypersensitivity as a proxy for changes in nucleosome occupancy/stability and chromatin accessibility) can be studied in order to test if chromatin

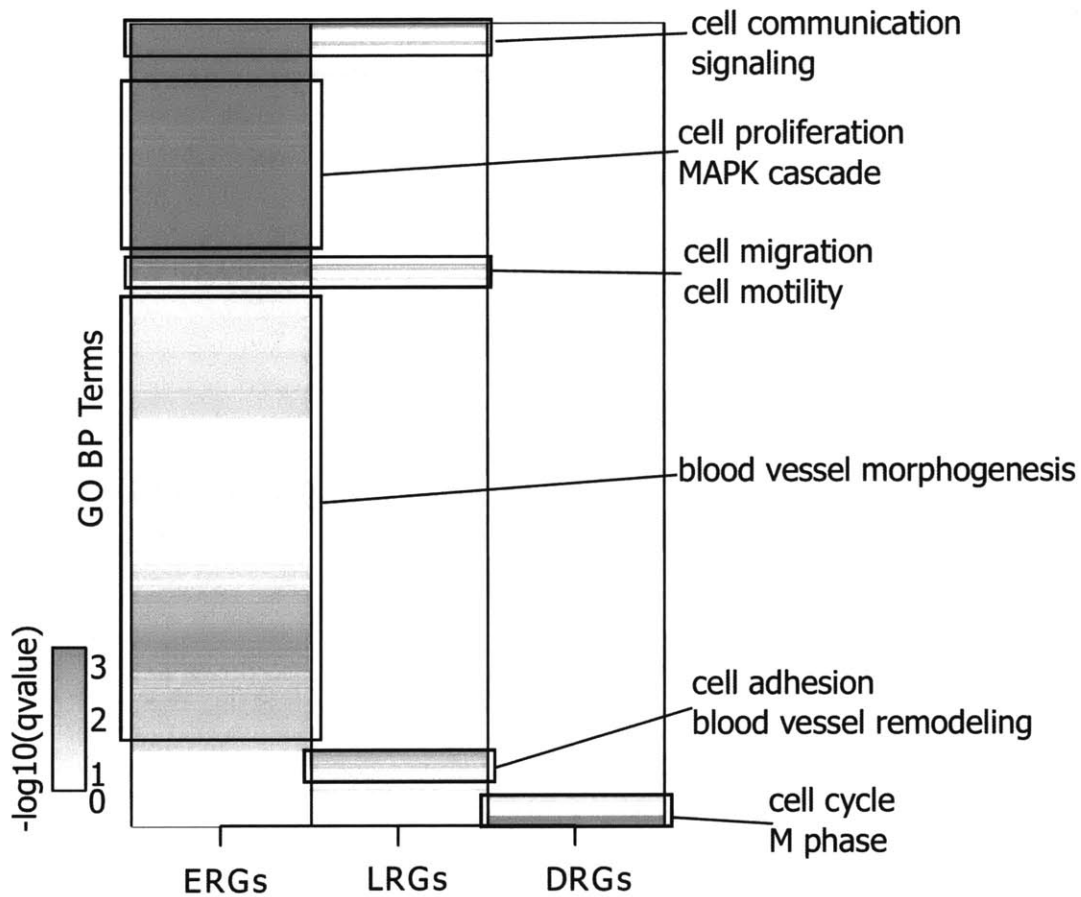


Figure 3-5: Enrichment of GO biological process terms significant in at least one of the gene sets (based on a 10% FDR cutoff). The enrichment is calculated relative to genes expressed within HUVECs (including NRGs). Overall, many of the key known angiogenesis functions (*e.g.*, proliferation, migration) appear to be within the ERGs, although they may be the more well-studied genes in angiogenesis to this point. LRGs are enriched in some similar terms as ERGs, but more uniquely enriched in functions relating to cell adhesion. DRGs are enriched for genes involved in the cell cycle. Terms highlighted reflect key patterns within the enclosed cluster, but other GO BP terms are enriched within each highlighted cluster as well.

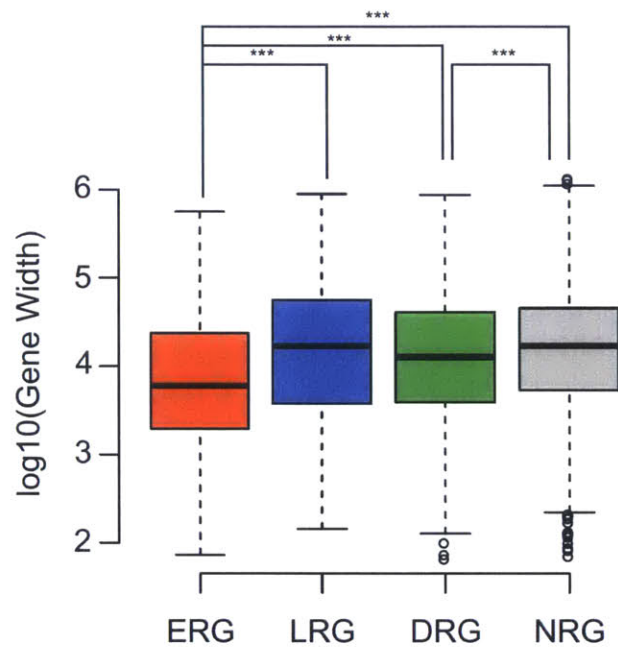


Figure 3-6: Comparison of the distribution of gene length in number of base pairs (exon plus intron) for each gene set. Overall, ERGs are significantly smaller than the other gene sets. The other three gene sets, except between DRGs and NRG, are generally of similar size. (Mann-Whitney U test, only significant comparisons are shown. \*\*\*  $p < .001$ )



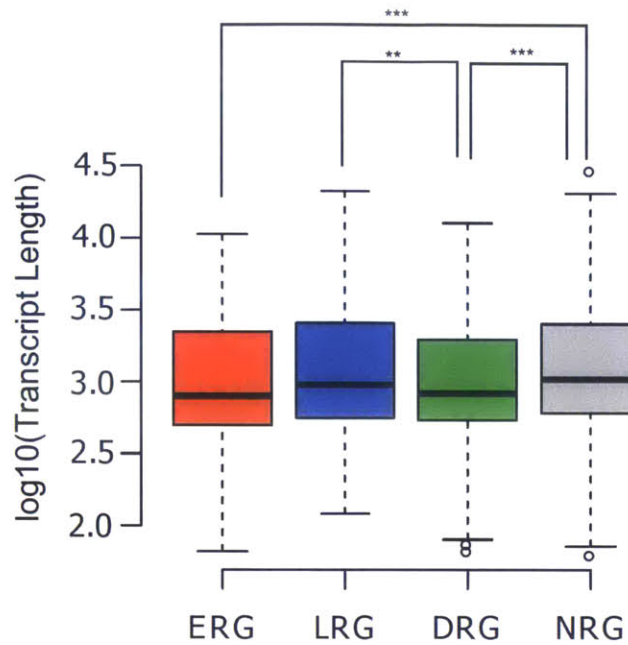


Figure 3-7: Comparison of the distribution of transcript length in base pairs (exon only) between each gene set. Although there are a few statistically significant differences between the gene sets, not all the differences seen at the gene length level are seen at the transcript level. (Comparisons performed by the Mann-Whitney U test, and only significant comparisons are highlighted. \*\*\*  $p < .001$ )

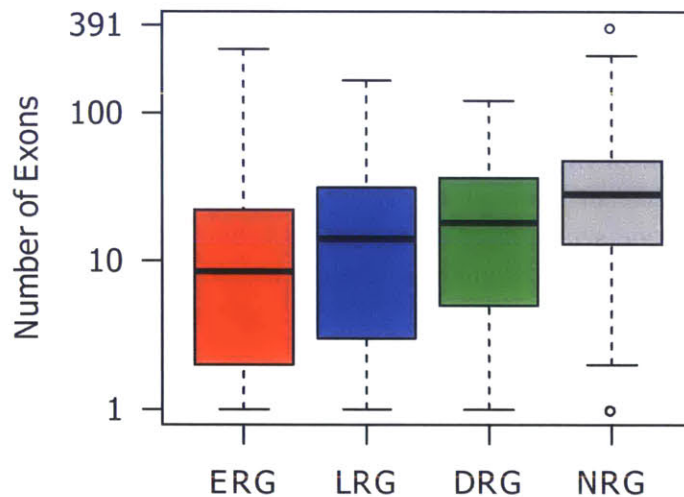


Figure 3-8: Boxplot of the exon count of the four gene sets (based on the ENCODE annotation). The number of exons per gene was determined by the total number of unique exons assigned to a particular gene. NRGs had the largest number of exons per gene on average. Consistent with their smaller size (see Figure 3-6), ERGs had the lowest number of exons on average per gene.

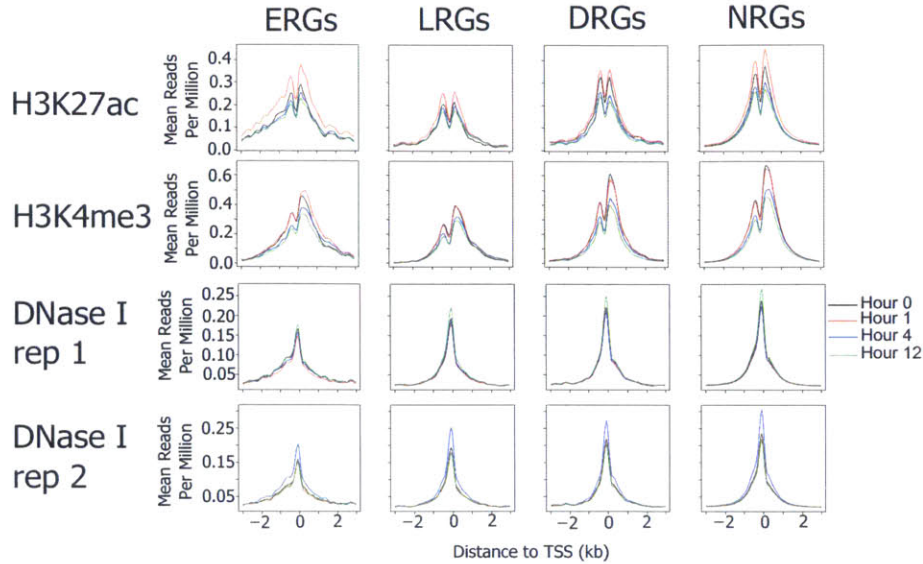


Figure 3-9: Average signal of H3K27ac, H3K4me3, and DNase I hypersensitivity of each gene set across the time course. In H3K27ac, there was temporal variation in ChIP-seq signal levels but non-specific to any particular gene set. It appears that H3K27ac generally increases at Hour 1 before falling. H3K4me3 had a similar pattern to H3K27ac. Comparatively, there was no consistent variation in DNase I hypersensitivity signal. Although there were non-specific dynamic changes, the enrichment of each of these studied features appeared to vary on average across each gene set.

remodeling at TSS is possibly necessary for completing angiogenesis.

All observed changes TSS for these three chromatin marks was surprisingly non-specific (see Figure 3-9). There were dynamic changes in both TSS H3K27ac and H3K4me3 on average. However, these changes occurred across gene sets and irrespective of their expression timing. Both histone marks generally increased on average at Hour 1 (more so for H3K27ac) and decreased over the rest of the time course. Comparatively, the DNase I hypersensitivity profile at TSS of each gene set was rather stable over time (see Figure 3-9). This suggested that changes in nucleosome positioning at TSS were not required on average to regulate VEGFA-responsive gene expression changes, possibly because human TSSs and promoters tend to be DNase I hypersensitive across most cell types[73]. Hence, although chromatin remodeling may occur in response to VEGFA, changes in gene expression appear to be uncorrelated on average with these three factors associated with expressed TSSs.

While the chromatin dynamics at TSS did not relate to gene expression change, the baseline average signal of each factor varied across each gene set (see Figure 3-9). NRGs had the strongest average signal of all four gene sets. ERGs had the weakest average DNase I hypersensitivity signal of all the four clusters but the nucleosomes near the TSS of ERGs had had strong H3K27ac and H3K4me3 signal (see Figure 3-9). Curiously, LRGs had a higher DNase I hypersensitivity signal, but weaker H3K27ac and H3K4me3 than NRGs. Although chromatin dynamic may not differentiate these VEGFA-responsive genes, the



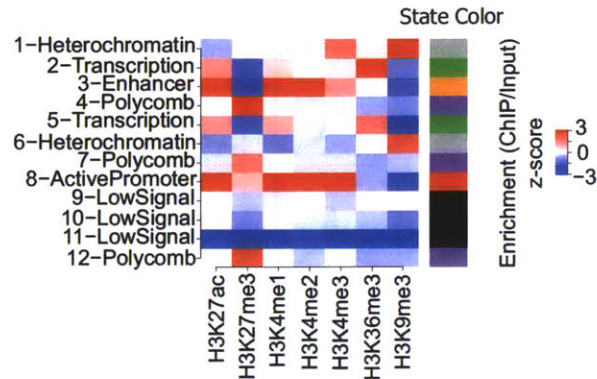


Figure 3-10: The chromatin state model produced from training a HMM on the six HUVEC data sets (each of the four time points and the two ENCODE HUVEC replicates[50]) for the 7 histone modifications listed here. The matrix reflects the mean signal of each histone modification per chromatin state.

chromatin signature at baseline may differentiate the temporal activity of different genes in HUVECs.

### 3.3.4 The Broad Chromatin Landscape is Stable under VEGFA Starvation and Stimulation

Using the additional histone modification ChIP-seq data (see Table 3.1), the HUVEC epigenome at each time could be annotated into chromatin states. A common method of segmenting and annotating a cell's epigenome are Hidden Markov Models (HMMs). Briefly, HMMs identify significant co-occurrences within the histone modification signal profiles that are predictive of the underlying chromatin state. This strategy has been used effectively in several studies[49, 50, 180]. In order to call the chromatin states at each time point, a HMM was trained on the smoothed, continuous signal profile of each of the 7 histone modifications (see Methods). In order to maximize biological relevance, the 12 state model was chosen because it appeared to balance segmenting the genome while having interpretable combinations of histone modifications within chromatin states (see Figure 3-10).

After calling the chromatin states in each sample (see Methods), comparing the chromatin state distribution between each sample showed little variation (see Figure 3-11). During the time course, the epigenome was rather stable to VEGFA stimulation. Hence, this suggests that most of the necessary regulatory elements for angiogenesis are active upon differentiation into HUVECs. It was interesting that the chromatin states distribution between the HUVEC time series and the ENCODE HUVEC samples was consistent. Although changes in growth conditions can cause changes in heterochromatin [128], the period of starvation prior to VEGFA stimulation did not appear to induce any such changes here.

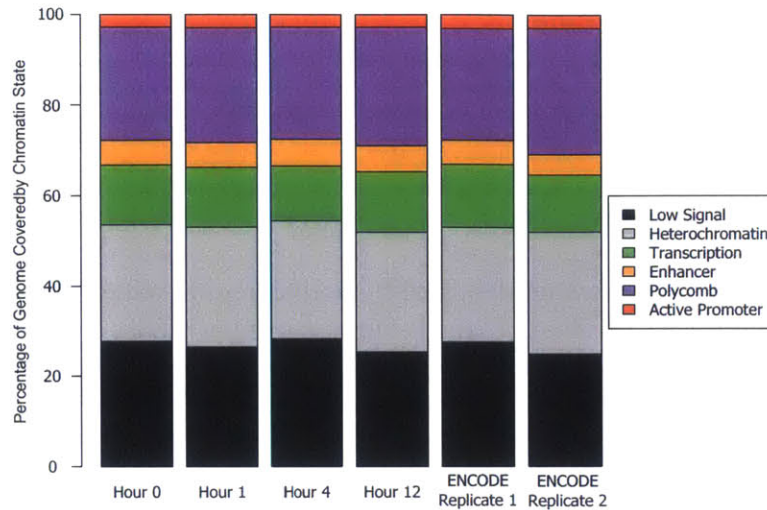


Figure 3-11: When grouping by chromatin state type per sample, the distribution of chromatin states was stable between samples including over time in response to VEGFA.

Type	Count
Total Enhancers	77,146
Super Enhancers	413
Latent Enhancers	1,610
Island Enhancers	1,583
Other Enhancers	71,907

Table 3.2: Frequency of each enhancer type within HUVECs. Note that the bottom four rows do not add up to the Total Enhancers row because super enhancers stitch nearby enhancers together, and each super enhancer is counted as one even if multiple independent enhancers were included to call one super enhancer.

### 3.3.5 The Enhancer Landscape during VEGFA Response

The chromatin state model identified candidate enhancer regions by identifying co-enrichment of H3K27ac and H3K4me1[58, 82] within one chromatin state. This enhancer chromatin state (see Figure 3-10) was used to identify the union of all candidate HUVEC enhancers within the time course. This generated 77,146 distinct candidate enhancer regions throughout the HUVEC genome (see Table 3.2). Using tethered chromatin capture (TCC) data for HUVECs grown in VEGFA[181] (where TCC is a Hi-C variant[182]), just under 50% of these enhancers had at least one long-range interaction, supporting many of these candidate enhancers are active, although it is likely that more of these enhancer are involved in long-range interactions but cannot be seen with the resolution of the TCC data. Overall, the chromatin state analysis allowed systematic identification of HUVEC candidate enhancers.

Super enhancers are strong, broad H3K27ac regions that are bound by a host of tran-



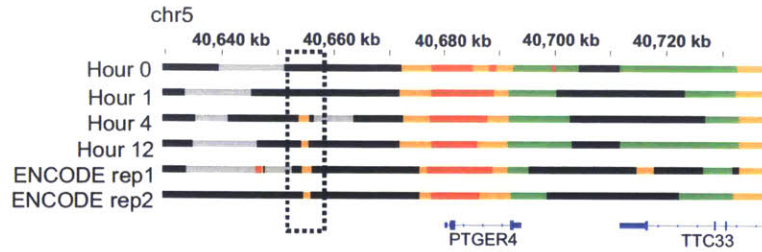


Figure 3-12: Latent enhancers were identified by finding regions where the enhancer chromatin state was not present at Hours 0 and 1 but appeared at Hours 4 and 12 (see dashed box). Several of these latent enhancers were active in the ENCODE HUVEC data set.

scription factors and regulate key cell identity genes[116]. Using the candidate HUVEC enhancers, enhancers within 12.5 kb were stitched together (see Methods). After stitching nearby enhancers, super enhancers were called per time point based on the cumulative H3K27ac signal as previously described[116]. The union of super enhancers across the 4 time points generated 413 HUVEC super enhancers (see Table 3.2).

While the epigenome was generally stable in response to VEGFA, there were multiple latent enhancers[85] activated. Latent enhancers were defined by identifying regions missing an enhancer chromatin state at Hours 0 and 1 but gained it at Hours 4 and 12 (see Methods and Figure 3-12), consistent with its previous definition[85]. There were 1,583 latent enhancers (see Table 3.2). Some latent enhancers were active in the ENCODE HUVEC data (see Figure 3-12), suggesting they are true enhancers. If the ENCODE HUVEC cells are representative of the pre-stimulation HUVECs (since they are grown in a similar conditions pre-starvation[150], see Figure 2-1), then these latent enhancers that were active in ENCODE HUVECs rapidly disappeared within the 12 hour starvation period prior to VEGFA treatment, which was faster than previously observed latent enhancers[85].

Curiously, at several locations across the genome there were enhancers surrounded by either H3K9me3 or H3K27me3 enriched regions without a nearby genes (see Figure 3-13). These 1,583 sites were termed *island enhancers* (see Table ??), since they were active regions within a sea of repressed chromatin. These regions tended to be active in the ENCODE HUVEC data and throughout the entire VEGFA time course (see Figure 3-13).

Surprisingly, p300 did not bind either latent or island enhancers throughout the time course despite H3K27ac enrichment (see Figure 3-14). Comparatively, p300 strongly bound super enhancers (see Figure 3-14), especially at Hour 1. It was especially surprising p300 did not bind latent enhancers given its histone acetyltransferase activity, although with the peaking of p300 binding at Hour 1[150]. Surprisingly, these three enhancer types were sensitive to C646 pre-treatment, where C646 inhibits p300 acetyltransferase activity[164]. Both super enhancers and island enhancers lost H3K27ac (see Figure ??), even though p300 did not bind island enhancers. Most surprisingly, latent enhancers gained H3K27ac enrichment at Hour 0 under C646 pre-treatment and maintained it throughout the time course (see Figure 3-15), which suggested p300 indirectly block latent enhancer activation.

Super, latent and island enhancers were all involved in long-range interactions[181]

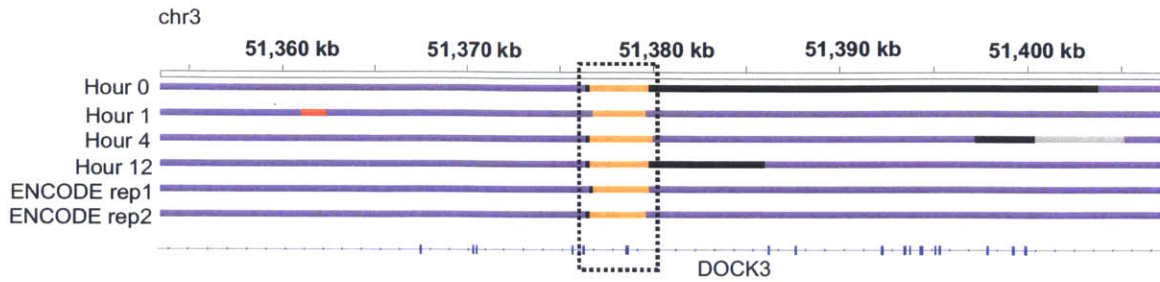


Figure 3-13: There are enhancer chromatin states that are active throughout the time course but are surrounded by heterochromatin (Polycomb/H3K27me3-repressed chromatin, in this case). An example island enhancer is within the dashed box, and this example is also an active enhancer within the ENCODE HUVEC data set also surrounded by heterochromatin.

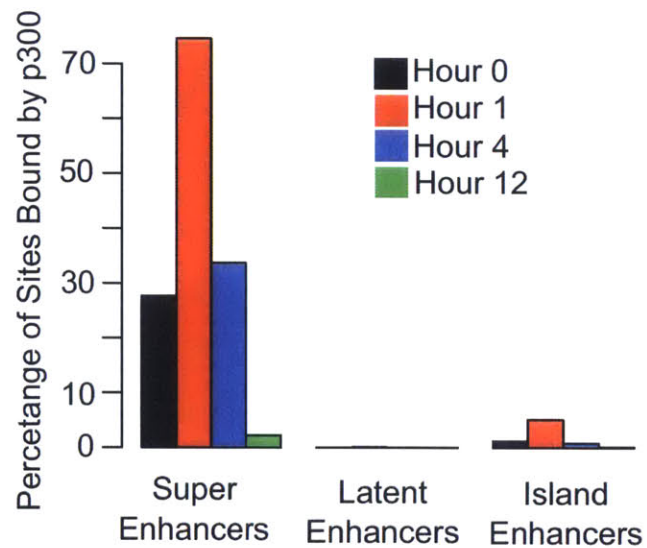


Figure 3-14: The percentage of super, latent, and island enhancers bound by p300 over time showed that p300 predominately binds super enhancers among these three enhancer sets.

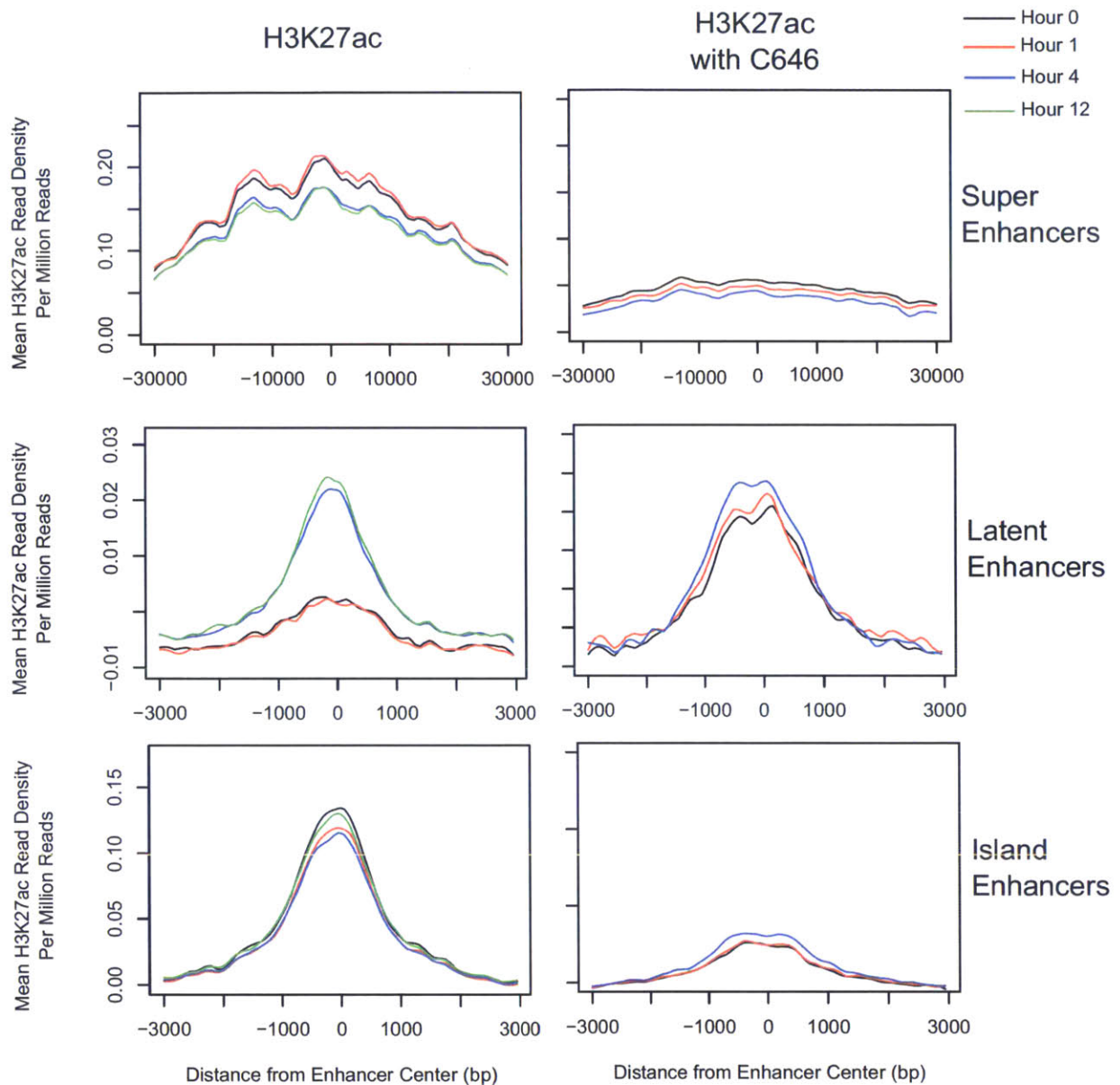


Figure 3-15: H3K27ac profiles over the VEGFA time course show that both super enhancers and island enhancers have strong H3K27ac throughout the time course while latent enhancers, as defined, gain H3K27ac at Hour 4 and 12. When the HUVECs are pre-treated with C646 before VEGFA stimulation, this changed temporal H3K27ac enrichment at all these enhancer types despite p300 only binding super enhancers (see Figure ??). Inhibiting p300 acetyltransferase activity with C646 suppressed H3K27ac at super and island enhancers but activated latent enhancers.



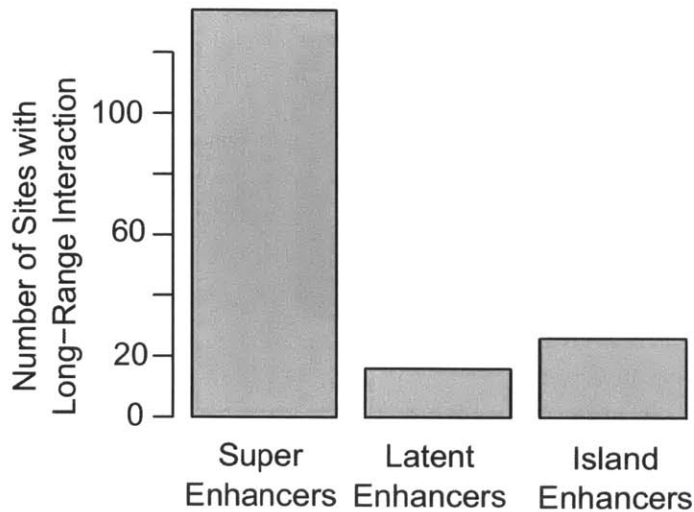


Figure 3-16: Using the long-range interaction data, the number of each type of enhancer was at one end of a long range interaction where a VEGFA responsive gene (ERGs, LRGs, and DRGs combined) was on the other end are counted here. Super enhancers had the most gene targets, but both island and latent enhancers did target these responsive genes.

with a promoter of a VEGFA-responsive gene (see Figure 3-16). Using the union of ERGs, LRGs, and DRGs, each enhancer type had a subset of individual enhancers involved in a chromatin loop targeting a VEGFA-responsive gene, suggesting that each enhancer type is important in angiogenesis. However, super enhancers predominately were involved in these long-range interactions compared to island and latent enhancers (see Figure 3-16).

Finally, the previously identified H3K27ac variant sites[150] were mapped onto HUVEC enhancers and expressed promoters to understand the distribution of VEGFA-responsive site (see Figure 3-17). About 20% of each set of H3K2ac variant sites mapped to super enhancers, suggesting super enhancers can dynamically regulate gene expression in signal-response. Virtually no H3K2ac variant sites mapped to island and latent enhancers, which is consistent with the fact that p300 did not bind these enhancer types and all the H3K27ac variant site are within 2 kb of a p300 binding site[150]. Otherwise, the majority of all H3K27ac variant sites were within typical HUVEC promoters and enhancers. Hence, the majority of VEGFA-responsive enhancers lie within non-super enhancers (see Figure 3-17) despite the key regulatory role super enhancers play in a cell. Hence, this suggests that typical enhancers and promoters may primarily regulate VEGFA-response.

### 3.3.6 Comparison of Active Regulatory Regions during VEGFA Response and across Endothelial Cell Subtypes

Although p300-based H3K27ac variant sites had no detectable change in DNase I hypersensitivity in response to VEGFA[150] (see Figure 2-14), this was small portion of active regulatory elements in HUVECS. DNase I hypersensitive peaks in each time point were called using HotSpot[183] and mapped to HUVEC enhancers (see Methods). Enhancer-mapped

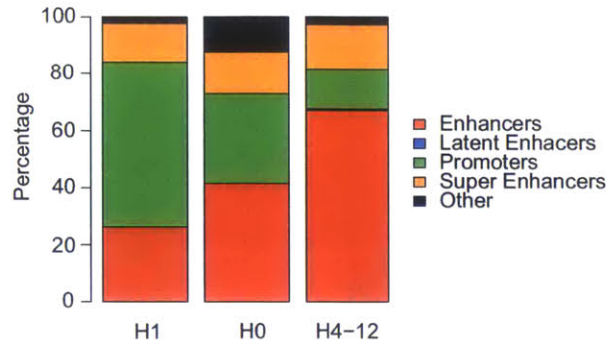


Figure 3-17: Within each H3K27ac variant cluster, a large fraction of sites mapped to super enhancers. Interestingly, the vast majority of variant H3K27ac sites within each cluster mapped to typical promoters or enhancers.

DNase I hypersensitive peaks enhancers were clustered (see Figure 3-18) and the majority of these DNase I hypersensitive sites were stable in response to VEGFA. active throughout the time course. But a number of peaks were responsive to VEGFA (see Figure 3-18), suggesting that outside of variant H3K27ac sites many enhancers undergo chromatin remodeling in response to VEGFA, possibly with only a change in DNase I hypersensitivity.

Endothelial cells are phenotypically diverse depending on their location in the body[1], and the regulation that drives these differences can be assessed with differences in DNase I hypersensitivity. Recently, the ENCODE consortium profiled multiple different endothelial cell types using DNase-seq[73], which can be compared against the DNase I hypersensitivity within the VEGFA time course to identify regulatory differences between endothelial cell types. DNase I hypersensitivity peaks from the time series data, ENCODE endothelial cell types and ENCODE GM12878 (as an comparison against a cell type that also arises from the mesoderm) were clustered using k-means clustering and broken down into 10 clusters (see Methods and Figure 3-19). In cluster 10 there was a high degree of shared DNase I hypersensitive peaks across all endothelial cells, but other clusters showed a high variability across all samples (see Figure 3-19). Notably, clusters 1 and 8 were mostly enriched for sites in HMBEC and HPAEC, respectively. HMBEC are endothelial cells from the brain vasculature, suggesting this cluster may contribute towards regulating the blood-brain barrier[1]. HPAEC are arterial endothelial cells, suggesting arterial endothelial cells have a very unique set of regulatory sites from these other vascular endothelial cells[1]. Despite the diversity in open chromatin across these cell types, overall cluster 10 suggests a highly shared set of candidate regulatory elements that may generally encode the regulation of angiogenesis among endothelial cells in general.

### 3.3.7 Genetic Variation within VEGFA Responsive Sites

Several studies have identified common DNA variants linked with various human diseases or traits through the use of genome-wide association studies[184] (GWAS). Since the majority of GWAS SNPs fall outside coding regions[185], mapping GWAS hits to variant



Figure 3-18: When mapping DNase I hypersensitive peaks at HUVEC enhancers over the time course, the majority of peaks were stable. Yet, there were a substantial number of peaks responsive to VEGFA, suggesting enhancers undergo chromatin remodeling outside of H3K27ac variant sites.

H3K27ac sites[150], may suggest whether a GWAS SNP contributes towards the regulation of a disease or trait[185–187]. Mapping the NCBI GWAS SNP catalog onto the H3K27ac variant sites revealed a number of GWAS SNPs fell within these regions. The count of the number of overlapping SNPs with H3K27ac variant sites are listed in Table A.1. Notably, several likely endothelial-related GWAS terms had SNPs fall within these regulatory elements, such as “Coronary artery calcification” and “Retinopathy in non-diabetics”. Additionally, other diseases with an implicated angiogenesis component had GWAS SNPs in diseases falling within H3K27ac variant sites[1], such as SNPs for Alzheimer’s Disease.

### 3.3.8 Multiple Dynamic Changes in Transcription Factor Co-binding Upon VEGFA Stimulation

Transcription factor binding drives changes in gene expression[14], so analyzing VEGFA-responsive transcription factor binding changes should shed light into what factors regulate VEGFA-responsive genes. For the following assayed transcription factors, p300, MYC, ETS1, GATA2, NICD, RBPJ, FLI, JUN, ERG and Pol II, ChIP-seq peaks were called for each per time point to identify candidate binding sites (see Methods). However since only one replicate was generated for each transcription factor per time point, and there can be a high false positive rate in transcription factor peaks calls between replicates[188]. In order to filter for peaks that are likely true binding events, transcription factor peaks within the same time point were overlapped. Regions where called peaks from at least two different transcription factors overlapped in the same point were kept as candidate *cis*



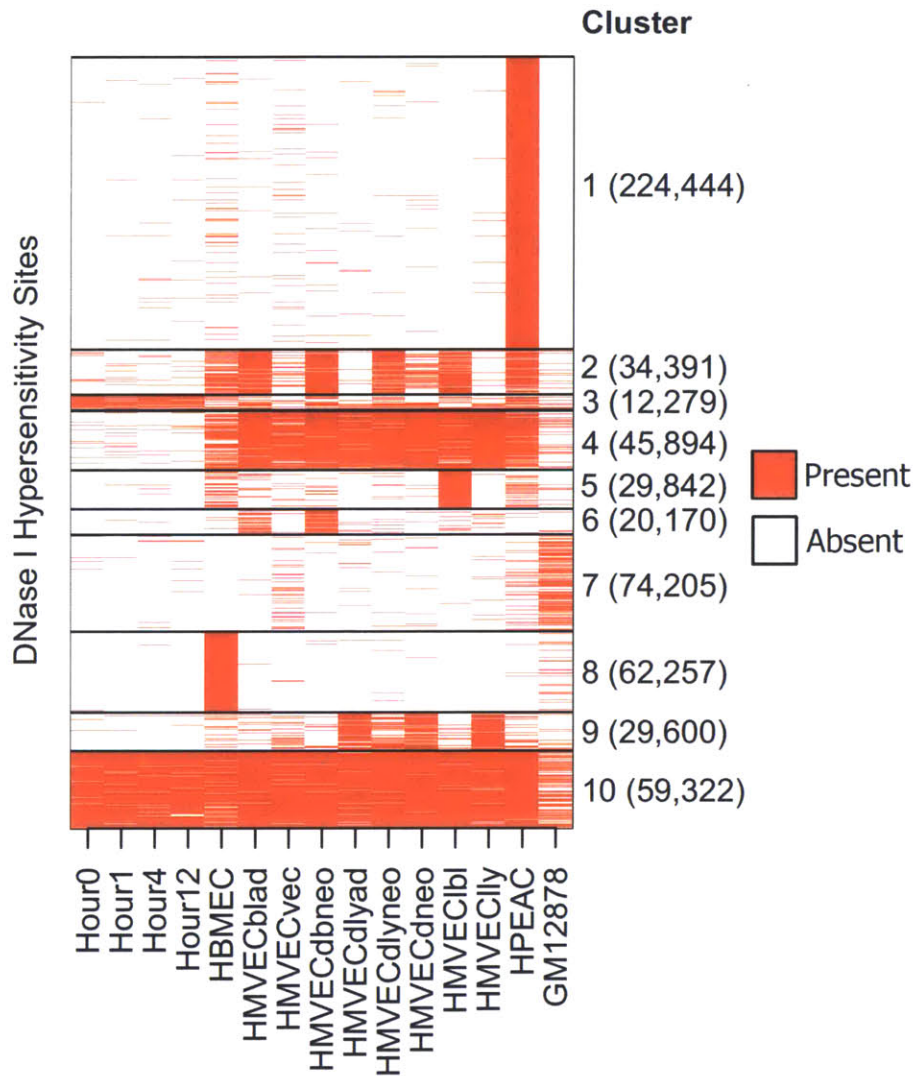


Figure 3-19: Mapping of DNase I hypersensitive peaks across multiple endothelial cells and GM12878 (a non-endothelial, mesodermal cell type). Using k-means clustering to divide the sites into 10 clusters (each cluster is listed on the right with the number of DHS peaks within it), there was a wide diversity of open chromatin regions across the different endothelial cell types, and independent of many regulatory regions found in the mesoderm derived GM12878. Within cluster 10, there was a high degree of shared DNase I hypersensitivity peaks across the time series and other endothelial cell types, supporting a common regulatory landscape. But there were peaks associated with specific endothelial cell types that could contribute towards differences in endothelial cell function. For example, cluster 8 were sites specific to HBVEC (human brain vascular endothelial cell), which are endothelial cells that contribute to the blood-brain barrier.

regulatory regions (see Methods). This overlap process generated 60,511 candidate *cis* regulatory regions. Once determining the candidate *cis* regulatory regions, all transcription factor peaks in any time point that overlapped a candidate *cis* regulatory region were kept, even if only one peak was found within a time point (see Methods). In order to focus on distal regions, candidate *cis* regulatory regions within 500 bp of an annotated TSS were removed. While this does not cover the full binding profile of each transcription factor per time point, it allows the identification of interesting VEGFA-induced transcription factor co-binding patterns that may drive at least a subset of VEGFA responsive genes.

Self-organizing maps (SOMs) were used to identify patterns of temporal transcription factor binding in candidate *cis* regulatory regions. SOMs have been used successfully in clustering and finding patterns in DNase I hypersensitivity and transcription factor co-binding[73, 114]. Here, a variation of the standard SOM method was used to account for the multiple time points and to capture how transcription factor co-binding clusters change over time in response to VEGFA. The SOM map has multiple levels or layers, where each layer corresponds to transcription factor binding in one time point. This enables clustering jointly over all four time points while allowing the identification and analysis of transcription factor co-binding patterns within a time point (see Methods). After generating the SOM, there were a variety of stable and dynamic transcription factor co-binding patterns at candidate *cis* regulatory regions in response to VEGFA (see Figure 3-20).

ETS1, an important endothelial transcription factor[5], binds pervasively to all candidate *cis* regulatory regions. ETS1 strongly binds all SOM clusters throughout the time course (see Figure 3-21). Comparatively, the number of candidate *cis* regulatory regions bound by p300 binds average throughout the time course is rather narrow (see Figure 3-22). While most p300 sites are co-bound by ETS1 (see Figure 2-12), p300 only binds a small fraction of the sites that are bound by ETS1. Hence, this may suggest that ETS1 acts as a lineage committing factor for the HUVECs. Although p300 is thought to be a major co-activator especially when it comes to multiple transcription factors binding, it appears ETS1 may have a broader role in mediating co-binding in this system since much of the observed co-binding and changes in transcription factor binding occur outside regions where p300 is found (see Figure 3-20).

Other transcription factors had a narrower binding profile compared to ETS1. JUN, FLI, and ERG are critical transcription factors in developmental angiogenesis[5]. In the time course, the binding of these three factors changed (see Figure 3-23). These three factors tended to occupy a similar set of candidate *cis* regulatory regions (see Figure 3-23). Since JUN is part of the AP-1 complex, the co-occupancy of these three transcription factors suggested ERG and FLI can co-bind with AP-1. JUN bound more candidate *cis* regulatory regions than either ERG or FLI, suggesting that it has functions independent of ERG and FLI. Comparatively, the more restricted binding of ERG and JUN at candidate *cis* regulatory regions was surprising considering the broad binding profile of ETS1 (see Figure 3-21). ETS1, ERG, and FLI are from the ETS family of transcription factors and share the common core GGAA consensus motif[109], but somehow ERG and FLI are able to bind a more selective set of candidate *cis* regulatory regions.

Looking at RBPJ, a critical factor in the Notch signaling pathway[3], its binding rapidly expanded over time across candidate *cis* regulatory regions (see Figure 3-24). At baseline, RBPJ had little to no detectable binding at candidate *cis* regulatory regions. But VEGFA

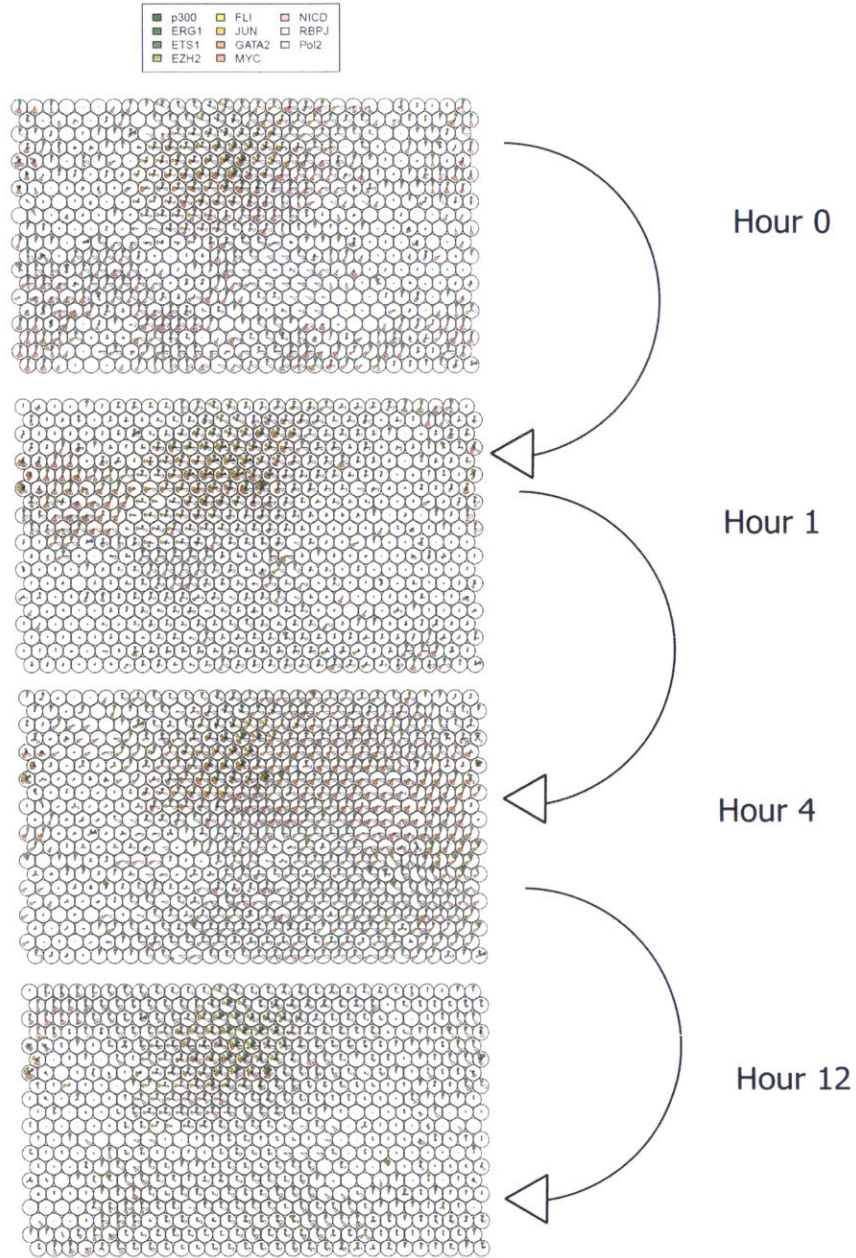


Figure 3-20: Using the 10 transcription factors assayed by ChIP-seq, the SOM clustering revealed multiple changes in transcription factor co-binding and location of binding over the four time points. The SOM map was passed the temporal transcription factor binding profile of 60,511 candidate *cis* regulatory regions and clustered the transcription factor binding patterns across the time course to identify common patterns of transcription factor binding change. The bars within each circle by color (as detailed in the key) reflect the average frequency of that transcription factor having a peak within all the sites clustered within the same circle (on a scale from 0% to 100%). It appeared the most transcription factor binding with these data sets increased heading into Hour 4 before tapering off in Hour 12.



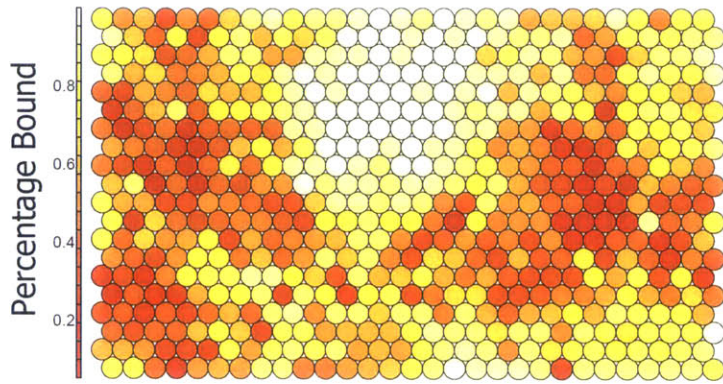


Figure 3-21: The mean occupancy of ETS1 binding within each cluster across the time course. Overall, ETS1 binds widely throughout all candidate *cis* regulatory regions.

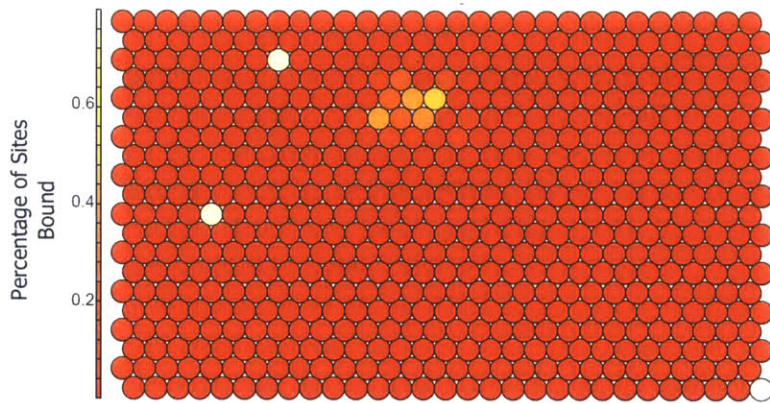


Figure 3-22: Across candidate *cis* regulatory regions, p300 binds a small set of clusters on average.

stimulation drove increased RBPJ binding, begging to a narrow set of sites at Hour 1 and rapidly expanding at Hours 4 and 12 (see Figure 3-24). Hence, it appeared that RBPJ binding was surprisingly expansive throughout candidate *cis* regulatory regions.

Surprisingly, GATA2 and MYC had two VEGFA-induced co-binding clusters. At Hour 0, these transcription factors co-occupied several SOM clusters but many other transcription factors also co-bound these clusters (black circles, see Figures 3-25 and 3-20). In response to VEGFA, these MYC and GATA2 rapidly co-bound a different set candidate *cis* regulatory regions (blue circles, see Figure 3-25). The VEGFA-induced co-occupancy of GATA2 and MYC occurred in only Hours 1 and 4. Surprisingly, the VEGFA-induced co-occupancy of GATA2 and MYC occurred at two different sets of clusters between Hours 1 and 4 (see Figure 3-25). Hence, it appeared that the VEGFA-induced cluster at Hour 1 was transient and the co-binding of GATA2 and MYC re-occurred at a new set of SOM clusters at Hour 4. As of this writing, the NCBI Gene database for both transcription factors has no documented protein-protein interaction between these two factors, suggesting this could be specific to VEGFA stimulation.

ERGs and LRGs appeared not to be driven by one particular set of co-binding transcription factors over time. When mapping each ERG and LRG to the nearest candidate *cis* regulatory region, neither ERGs nor LRGs tended to have a few common SOM clusters nearby. Both ERGs and LRGs mapped to multiple SOM clusters across the map (see Figure 3-26). ERGs and LRGs were mapped to clusters that included VEGFA-induced GATA and MYC co-binding to clusters with a large number of bound transcription factors over time (see Figure 3-20). This suggests that angiogenesis may be driven by multiple parallel pathways given the wide variety of transcription factors that appear to potentially drive upregulated VEGFA-responsive genes.

### 3.4 Discussion

This study sheds new light into underpinnings of transcriptional regulation of angiogenesis by analyzing VEGFA-induced gene expression, chromatin and transcription factor binding changes, complementing previous studies in developmental angiogenesis [5]. Analysis of VEGFA-responsive genes identified multiple temporal gene expression patterns, consistent with previous reports[150, 159], including many, previously unidentified non-coding RNA genes differentially expressed in concert with protein-coding genes. However, the function of almost all of these non-coding RNA genes is still unknown, which requires further in order to better understand how they regulate angiogenesis alongside of differentially expressed protein-coding genes. Further analysis of the gene structure of differentially expressed genes revealed a relationship between gene size and number of exons with the timing of gene expression change. This was particularly evident in ERGs that had a smaller gene size and few exons, suggesting that the smaller size could optimize rapid upregulation by allowing Pol II to finish transcribing faster.

Analysis of dynamic chromatin structure changes during VEGFA response revealed a complex set of changes. Broadly, chromatin state analysis suggested the HUVEC epigenome was rather stable, suggesting most regulatory elements necessary are already active. Yet, there were multiple latent enhancers that activated in response to VEGFA in addition to

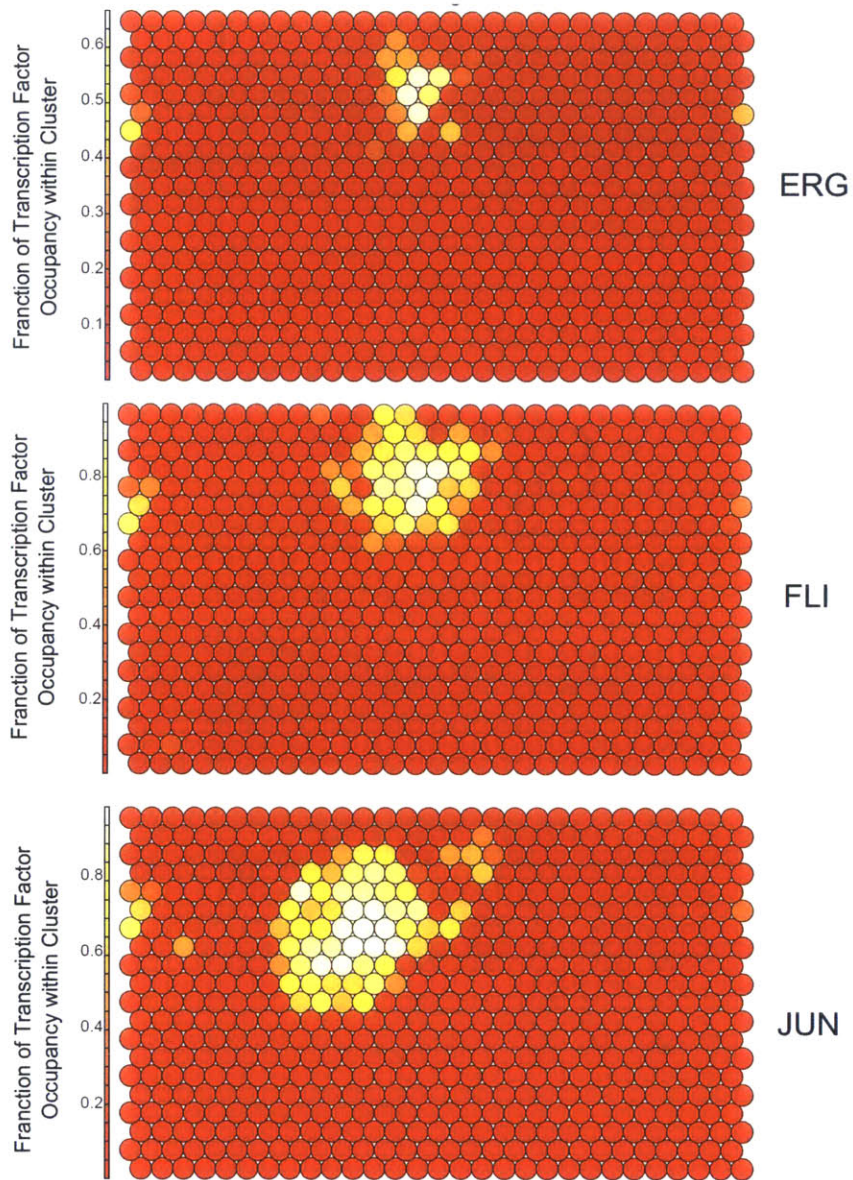


Figure 3-23: The average ERG, FLI, and JUN transcription factor occupancy within the SOM maps shows each factor has a much narrower binding profile compared to ETS1, especially ERG and FLI which are within the same ETS family of transcription factors as ETS1. Moreover, the binding of these three transcription factors has a strong co-occupancy with each other in several SOM clusters (upper middle).



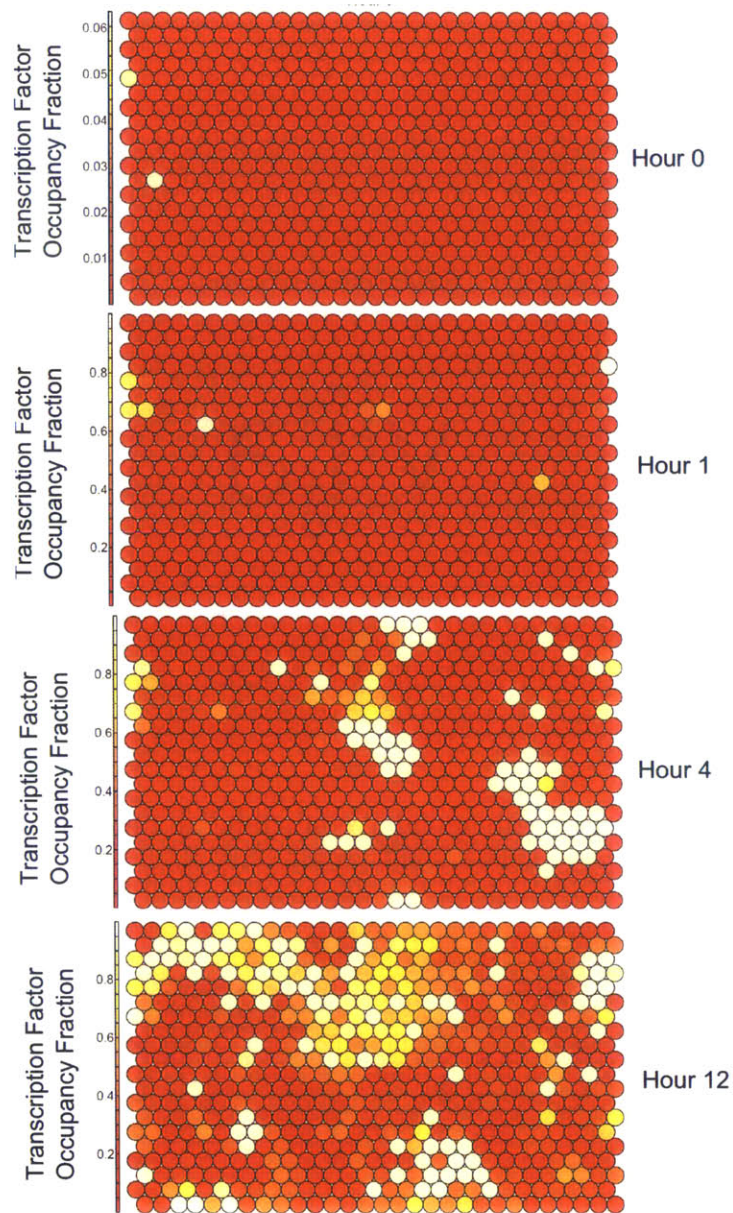


Figure 3-24: RBPJ binding increased throughout the VEGFA time course. Initially, RBPJ has little to no detectable binding at Hour 0 at candidate *cis* regulatory regions. In response to VEGFA, the binding profile of RBPJ dramatically increases throughout the time course, binding almost half of all clusters within the SOM by Hour 12.



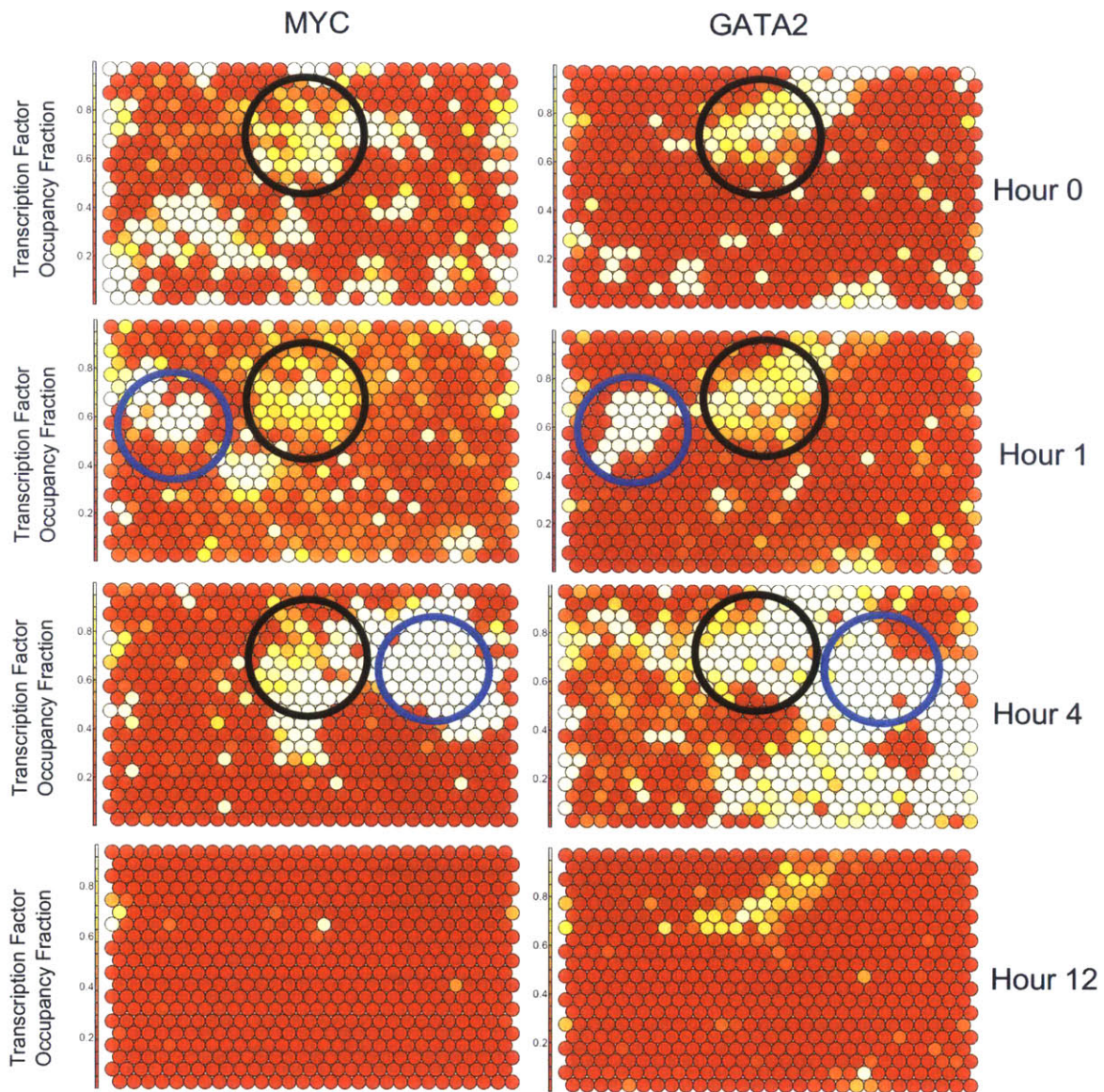


Figure 3-25: GATA2 and MYC have VEGFA-driven co-occupancy during the time course, especially at regions that lacked GATA2 and MYC binding pre-stimulus. In the black circle, MYC and GATA2 frequently co-occupy the same regions, although these regions within the SOM are co-occupied by a variety of other assayed transcription factors (see Figure 3-20). However, in the blue circle are regions that generally gained GATA2 and MYC binding at Hours 1 and 4 even though they lacked occupancy of these factors before VEGFA stimulation. Surprisingly, VEGFA-responsive co-occupancy of GATA2 in MYC is very dynamic since it binds one set of sites at Hour 1 and then a separate set of sites in Hour 4.

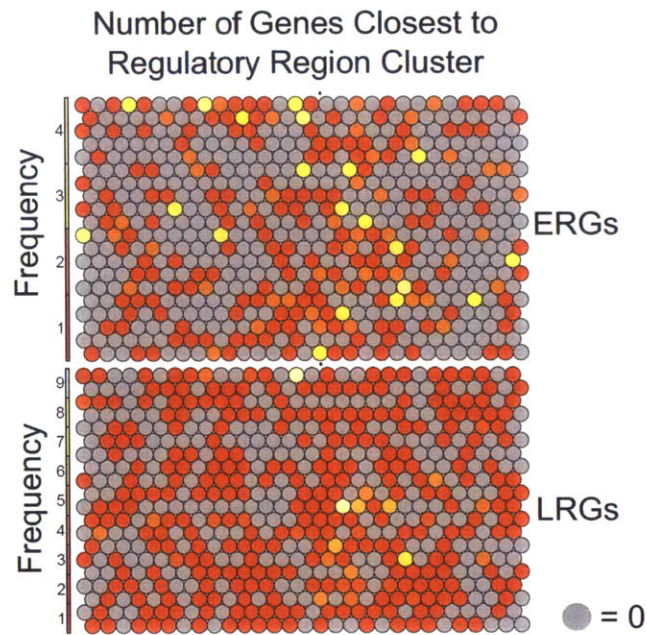


Figure 3-26: For the ERGs and LRGs, each gene within the set was assigned to the SOM cluster that had the candidate *cis* regulatory regions closest to the gene. The frequency that a responsive gene was mapped to each SOM cluster is shown within the figure for each gene set. Overall, no particular SOM cluster appeared specific to either ERGs or LRGs.



previously identified H3K27ac variant sites[150], suggesting that remodeling of histone modifications likely occurs at very specific loci. Additionally, HUVEC enhancers had a number of dynamic DNase I hypersensitivity peaks change over the time course, suggesting changes in nucleosome positioning away from dynamic H3K27ac sites that may enable time-specific transcription factor binding sites. Curiously, there was no specific chromatin remodeling change to TSS at VEGFA-responsive genes, suggesting that most chromatin remodeling that may help regulate changes in transcription occurs more distally. Finally, analysis of DNase I hypersensitivity peaks across endothelial cells revealed both a common set and type-specific set of regulatory elements. These differences could enable the better understanding of what angiogenesis-related regulatory elements are shared across endothelial cells.

Analysis of transcription factor binding changes revealed several new insights co-binding modules that regulate angiogenesis. Except for ETS1 that binds rather widely across all candidate *cis* regulatory regions across, many of the assayed transcription factors had a specific VEGFA-induced binding profile. One particular surprising finding was the VEGFA-induced GATA2 and MYC clusters at Hours 1 and Hour 4. There is no known GATA2 and MYC protein-protein interaction (according to the NCBI database), and this co-binding starts at one set of candidate *cis* regulatory regions at Hour 1 before moving a separate set of sites at Hour 4. Overall, these analyses shed light into the complex network driving angiogenesis and suggested a variety of transcription factor co-binding modules may regulate VEGFA-responsive genes.

## 3.5 Methods

### 3.5.1 Data Generation

All ChIP-seq, DNase-seq, RNA-seq, and related experiments were carried out in similar conditions as described in Zhang *et al.*[150]. C646 pre-treatment was carried out by using the same experimental setup as the normal VEGFA time course but pre-treating the HUVECs 30 minutes prior to VEGFA stimulation[150].

### 3.5.2 ChIP-seq Data Processing, ChIP-seq Peak Calling for Transcription Factors and Co-activators

Each ChIP-seq FASTQ file was aligned against human genome build hg19 using bowtie2[60] using standard parameters except for using the parameter -M 1 to minimize mapping of repetitive reads. Each transcription factor ChIP-seq data set were called with the *spp* R package[63] on R version 3.0.1. Peak calls on all transcription factor (or co-activator) data were made using the *find.binding.positions* function within the package for multiple FDR thresholds. The FDR threshold selected for each transcription factor depended on the quality of the ChIP-seq data. But the FDR threshold was never less than 15% and the same FDR threshold was used across all time points for the same transcription factor.

### 3.5.3 RNA-seq Data Processing and Gene Expression Quantification and Defining VEGFA Responsive Genes

RNA-seq data sets were mapped against human genome build hg19 using the GENCODE reference[178] with tophat2[189]. FPKM quantification was done with cufflinks2[173]. Gene quantification was filtered by the quality estimate in cufflinks2 if the software identified a quantification as “FAIL” at some time point, which the entire gene at each time point was removed from further analysis (removes < 1% of genes). To capture the differentially expressed genes, fold-change against Hour 0 was measured, taking all genes with at least a minimum 2-fold expression change, up or down, in at least one time point. Genes were removed if they never had an FPKM greater than 1 at any time point as well as genes with a FPKM of 0 at Hour 0 (since these tended to be false positives). To improve quality of differential expression calls, genes with less than 10 RNA-seq reads were discarded. Finally, all genes needed to have at least one Pol II ChIP-seq peak mapped to the GENCODE annotated gene region to be considered expressed. k-means clustering using averaging over 10 random starts and up to 50 iterations (to increase the robustness of the expression clusters). In clustering, the FC changes were capped within the [-3,3] range, to focus on the temporal patterns rather than the absolute fold change. Genes sets were categorized by their change in the time course.

For the purposes of this study, non-coding RNAs included all genes that were not identified as protein coding within the GENCODE annotation. This includes lncRNAs, antisense transcripts, processed transcripts, and pseudogenes.

### 3.5.4 GO BP Enrichment Analysis for Responsive Genes

Using the topGO Bioconductor R package[190], the GO biological process enrichment was calculated for the ERG, LRG, and DRG gene sets as described above. Non-protein coding genes were excluded from the analysis (as defined by the GENCODE annotation). Instead of using the full set of protein coding genes in GENCODE, only genes that were expressed in at least one time point were used for the background (that is, the union of all the protein-coding genes in the ERG, LRG, DRG, and NRG gene sets) to emphasize particular functions relative to what genes are normally active in normal endothelial function. Using this background, a p-value for each GO BP term was calculated using the Fisher test. These p-values per gene set were converted into q-values using the *qvalue* Bioconductor R package[191] using default parameters. The q-value cutoff for analysis was 10% FDR. Hierarchical clustering of the GO BP terms were only performed on those that passed the significance threshold in at least one of the three tested gene sets.

### 3.5.5 Chromatin State Calls using Hidden Markov Model Segmentation

In order to call chromatin states, a Hidden Markov Model was used. The HMM was implemented using the *RHmm* R package on R 3.0.1. A multi-dimensional Gaussian HMM model was used to model the enrichment distribution and interaction between the 7 chromatin marks used in the model. For each mark, the enrichment of each histone modification

was calculated across the genome in 200 bp windows, where enrichment was *ChIP/input* normalized to the library size of each. To normalize the range across histone modifications, each was converted into z-scores genome wide. The model was trained on the z-scores. The model was trained up to 100 iterations, and the initial parameterization of the model was generated from k-means clustering. The model was run from 8 to 15 states, choosing 12 as the best balance of segmentation and biological interpretability. After the model was trained, the chromatin states were called on the histone modification z-scores using the Viterbi algorithm.

### 3.5.6 Enhancer Analysis

HUVEC enhancers were identified through the annotated “Enhancer” chromatin state. The union of all enhancer chromatin states per time point was taken to be the set of candidate HUVEC enhancers. Long-range chromatin interaction analysis was based on called long-range interactions from TCC data for HUVEC as acquired from Kaikkonen *et al.*[181]

### 3.5.7 DNase I hypersensitivity analysis

DNase-seq data was used from previously published data sets: Zhang *et al.* for the time series and ENCODE for the other cell types. Reads were mapped in the same fashion as the ChIP-seq data. Peaks were called using HotSpot[183] with a 10% FDR cutoff. For all cell types except HPAEC, there were two replicates (all peaks were used from HPAEC since it only had one replicate), and only called overlapping peaks were used between the two replicates.

For the clustering the DNase I hypersensitivity peaks, all the peaks were merged into contiguous regions with a maximum 500 bp gap (in order to account for slight variations in the border between regulatory regions found). The overlapping sets were mapped back to this full set of regions. Then the matrix was clustered using k-means clustering with 10 groups. 10 groups via k-means cluster explained 70% of the observed variability, and marginal increase in variability explained with a higher number of k-means clusters started to rapidly decrease after 10 clusters.

### 3.5.8 Co-clustering of Transcription Factors

Using the ChIP-seq peak calls from each of the transcription factors, the peaks were overlapped across all time points in order to find the candidate *cis* regulatory regions allowing up to a 500 bp gap in order to accommodate possible transcription factors that may be bound at a candidate *cis* regulatory regions and help establish the co-binding pattern but were not assayed via ChIP-seq in this experiment. Reducing this gap parameter down to 0 bp did not have a significant impact on the overall result. Then within each time point, the binding of individual transcription factors was mapped back to this master set of regions. If a region did not have at least two different transcription factor binding sites within the *same* time point, it was discarded from further analysis.

The clustering using self-organizing maps was adapted from Xie *et al.*[114]. The self-organizing map implementation was from the *kohonen* R package. Instead of using the

standard *som* function, the *supersom* function was used instead with each layer being a separate time point. The function was fed a binary matrix per time point of transcription factor binding to a given region (if a particular transcription factor bound more than once of a region in the same time point, the extra binding sites were ignored).

# Chapter 4

## A Comprehensive Analysis of RNA Polymerase II Pausing Across Mammalian Cell Types

**Contributing Authors:** Daniel S. Day, Bing Zhang, Sean M. Stevens, Francesco Ferrari, Hojoong Kwak, Erica N. Larschan, Peter J. Park, and William T. Pu

**Contribution:** All the analyses within this chapter are my own work unless otherwise noted.

**Manuscript status:** Adapted from paper in submission

### 4.1 Abstract

RNA Polymerase II (Pol II) stably pauses before entering productive elongation of many genes. Although Pol II pausing has been implicated in regulating development and disease, its effects on mammalian gene expression and chromatin structure remain poorly understood. Here, we analyzed 280 genome-wide datasets, including 85 Pol II ChIP-seq experiments from 35 different murine and human samples, to gain new insights into the relationship between Pol II pausing and gene regulation. Across cell and tissue types, paused genes composed 60% of expressed genes, repeatedly associated with specific biological functions, and were enriched for genes recurrently mutated in human cancers. Paused genes also had lower cell-to-cell expression variability. Increased pausing had a non-linear effect on gene expression levels, in which moderately paused genes were overall more highly expressed than the most or least paused genes. Highest gene expression levels were often achieved through a novel pause release pathway driven by high Pol II initiation. Stimulus-responsive genes were overall less paused than non-responsive genes, and rapid gene activation was linked with conditional pausing release. Local sequence composition and chromatin structure near the transcription start site both influenced pausing, with divergent features between mammals and *Drosophila*. Most notably, pausing was positively correlated with H2A.Z promoter deposition and involved more distant chromatin regulation in mammals. Our results provide new research directions into how Pol II pausing contributes to mammalian gene regulation and will inform efforts for therapeutic

modulation of Pol II pausing.

## 4.2 Introduction

Prior to entering elongation of a subset of genes, initiated Pol II pauses just downstream of the transcription start site (TSS) by forming a stable complex containing a short, nascent RNA transcript[77, 136, 139, 145]. This promoter-proximal pausing of Pol II (“Pol II pausing”) is regulated by protein complexes, with negative elongation factor (NELF) promoting pausing and positive transcription elongation factor (P-TEFb) stimulating pausing release[77, 136]. While discovered at *Drosophila* heat shock genes[77], Pol II pausing occurs widely in many metazoan genomes and regulates diverse biological functions[117, 143, 192–194]. Despite the expanding literature about Pol II pausing, its functional role remains uncertain[77]. In particular, the relationship of Pol II pausing to the expression level and surrounding chromatin structure of a gene is poorly understood, especially in mammalian cells.

Here, the functional significance of Pol II pausing is investigated by analyzing a combination of published and new ChIP-seq (chromatin immunoprecipitation followed by high-throughput sequencing) datasets on the genome-wide distribution of Pol II in mammalian cells and tissues, in conjunction with related datasets on gene expression and chromatin landmarks. This comprehensive analysis reveals previously unexpected relationships between Pol II pausing, gene expression, chromatin features, and the pausing regulatory machinery.

## 4.3 Results

### 4.3.1 Characterization of Pol II pausing across multiple cell types

The breadth and strength of Pol II pausing was estimated from 64 human and 24 mouse mostly publically available Pol II ChIP-seq datasets spanning multiple cell lines and tissue types (Supplementary Table B.1). Pol II ChIP-seq data was used to estimate transcriptionally productive Pol II occupancy rather than other techniques such as GRO-seq[145] or PRO-seq[139] because ChIP-seq data was publically available for a much wider range of mammalian cell types while providing sufficient sensitivity to capture productive Pol II[147]. For each gene, we quantified pausing by calculating its *Pausing Index* (PI; also referred to in the literature as the *Traveling Ratio*)[77, 143, 144, 195], defined as the ratio of Pol II ChIP-seq sequence read density in the TSS region (TSSR, -50 to +300 bp around TSS) to the gene body density (TSS+300 bp to +3 kb past the transcriptional end site (TES); see Figures 4-1 and B-1 and Methods). The PI index measures the build-up of Pol II at the TSS, as a surrogate for estimating what frequency Pol II is found at the paused state at TSS for a particular gene. Since Pol II is often found in high density at the TSS in mammals[77, 145], the PI quantifies the relatively frequency of how often Pol II is likely found in the paused state at TSS rather than elongating at steady-state[77]. To minimize noise from genes with low transcriptional activity, genes with Pol II and H3K4me3 TSSR



density below threshold values were not assigned PIs (see Methods). For genes with multiple annotated TSSes with strong Pol II and/or H3K4me3 TSSR density, we designated the one with the strongest H3K4me3 signal as the gene’s primary TSS in order to capture the most active promoter (see Methods). Our PI estimates correlated well across biological replicates (Figure B-1), even when comparing across different Pol II antibodies. Furthermore, two independent markers of Pol II elongation, H3K36me3 and Pol II phosphorylated on serine 2 (Pol II pS2), strongly correlated with our ChIP-seq based gene body Pol II density estimates (Figure B-2), suggesting that we accurately quantified elongating Pol II.

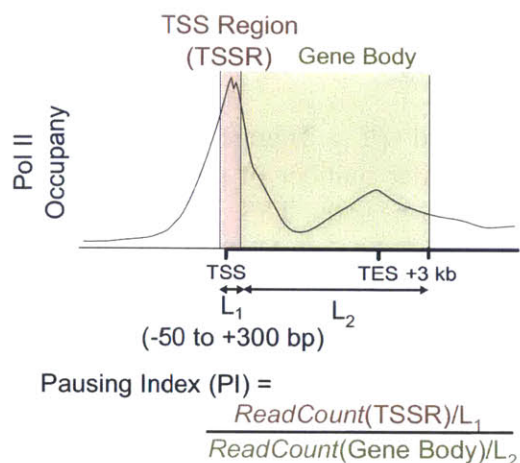


Figure 4-1: The definition of the “Pausing Index” (PI) to estimate how paused Pol II is at a gene’s TSS. The PI is the ratio between the Pol II occupancy at the gene’s TSS (-50 bp to +300 bp around TSS) and the gene’s gene body (+300 bp from TSS to +3 kb past the gene’s TES). Pol II occupancy is measured by the Pol II ChIP-seq read count in each region, normalized by the library size and width of the region. As the PI increases, this is to be interpreted that Pol II at TSS is “increasingly paused”. For more, see Section 4.5.2.

*Paused genes* were identified in each analyzed cell type based on having a significant build-up of Pol II at the TSS relative to the gene body (genes with a minimum two-fold TSSR density over gene body density, *i.e.*,  $PI > 2$ ; see Methods) and accounted for about 33% and 46% of RefSeq annotated genes across each human and mouse sample, respectively (see Figure 4-2, a representative sample of cell types studied). The frequency of paused genes per cell type was surprisingly consistent across different cell states (*e.g.*, embryonic, adult, or cancer). Concurrently, the majority of paused genes were paused across at least 75% of human or mouse samples (see Figure 4-2). This suggested that for many genes the pausing of Pol II at TSS may not be dependent on the particular cell type.

The function enrichment for paused genes was identified by analyzing enrichment within the Gene Ontology (GO) terms for biological functions that within genes in the top and bottom quartile ordered by PI in each sample. We found that the most and least paused genes were recurrently enriched for several biological function terms across most cell types (see Figures 4-3 and B-3). Genes on the top of the PI distribution were enriched for GO terms involving cellular metabolism, DNA repair, protein localization and cell cycle, whereas genes towards the bottom of the PI distribution were enriched for de-

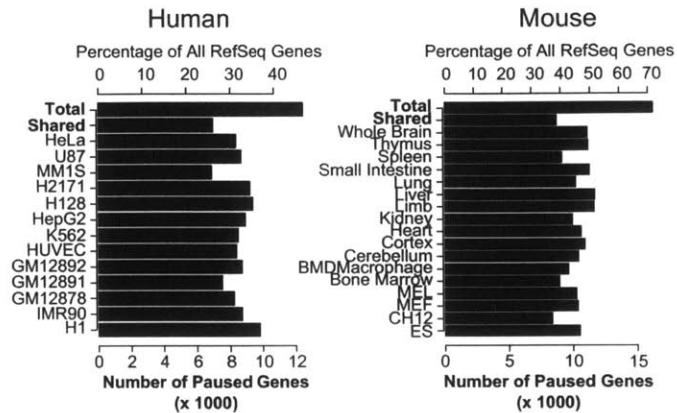


Figure 4-2: The number of paused ( $PI > 2$ ) genes per cell type in human and mouse revealed most cell types had a similar number of paused genes, where pausing suggested the balance of observed Pol II fell near the TSS rather than across the gene body. Given the similar fraction of paused genes per sample, there was a high level of “shared” paused genes across cell types, where a paused gene was considered “shared” if it was paused in at least 75% of cell types for that organism.

developmental, apoptosis and cell signaling terms (see Figures 4-3 and B-3). These GO term enrichments were maintained when we categorized genes into strongly or weakly paused categories by their median PI rank across all samples (see Methods and Figures 4-3 and B-3). While some of these GO terms (*e.g.*, cell cycle) have been previously identified as enriched in mouse stem cells[143], our study identified additional GO terms enriched amongst the most or least paused genes, and demonstrate that these functional enrichments hold across multiple cell types.

Surprisingly, strongly paused genes tended to be targets of recurrent somatic mutations that lead to cancer. Recently it was reported a set of about 300 human genes that are significantly, recurrently mutated in one or more cancer types[196]. Paused genes across cell types significantly overlapped genes that were recurrently the target of somatic mutations in cancer. Both paused and strongly paused genes were enriched for genes in this reported gene set ( $oddsratio > 2$ ,  $p < 0.001$ , Fisher exact test; see Figures 4-4 and ref-fig:pausingsuppl3). While it is unclear whether paused Pol II influences the occurrence of somatic mutations, nonetheless strongly paused genes appear to be functionally beneficial towards tumor progression. Collectively, our analyses show that Pol II pausing is widespread and regulates genes in similar functional categories in different studies and is involved in important physiological and pathological processes.

Given that many genes were recurrently paused across cell types, we investigated the correlation of primary DNA sequence characteristics with Pol II pausing. In *Drosophila* multiple motifs, such as the TATA box, associate with paused genes [139, 193, 197, 198], and in humans the TSS of paused genes in IMR90 qualitatively tended to be within 2 kb of an annotated CpG island[145]. To explore whether these relationships were generally valid for human and mouse cell types, we quantitatively compared the promoter ( $\pm 500$

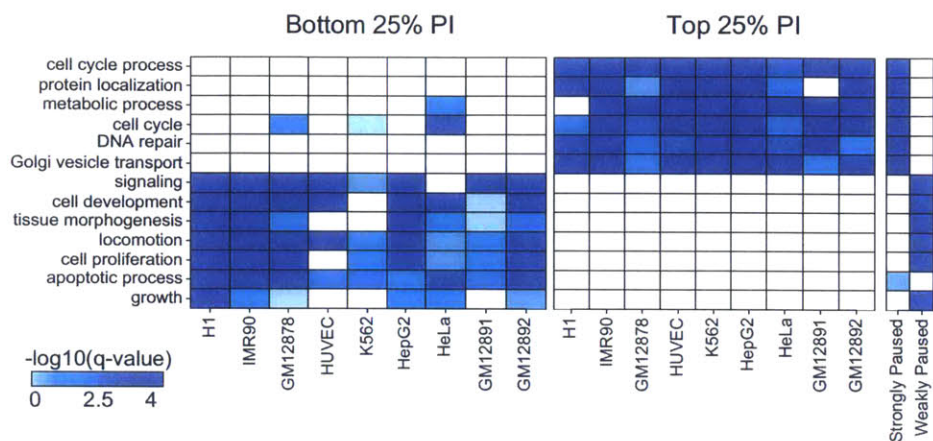


Figure 4-3: For the subset of cell types listed here, enrichment for GO biological processes were calculated for genes on the top and bottom 25% of genes within the cell type. High and lowly paused genes were based on the median PI rank across all analyzed samples (whether its median rank was in the top or bottom half, respectively, of all genes with an assigned PI).

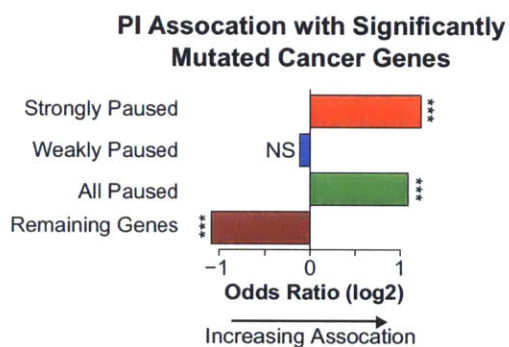


Figure 4-4: When taking the set of paused and non-paused genes from across all analyzed cell types and overlapping them with a list of recurrently mutated cancer genes across multiple cancers[196], there was a strong association with highly paused genes (and paused genes in general) with this set of about 300 cancer genes. (\*\*\*)  $p < .001$ , NS = not significant, Fisher exact test)



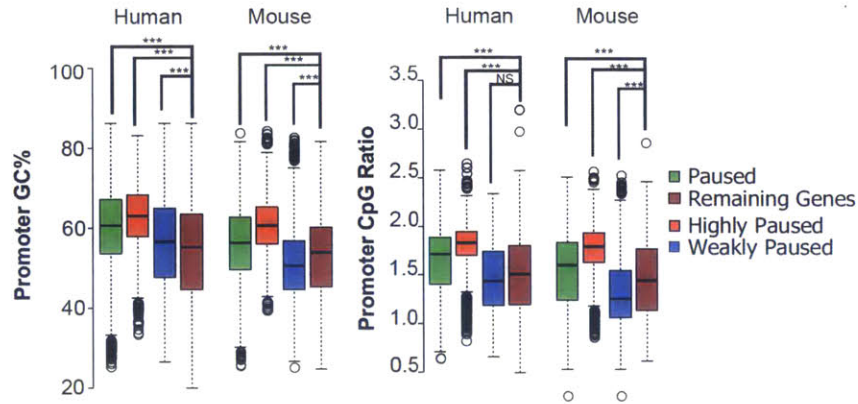


Figure 4-5: A quantitative assessment of the GC and CpG content at paused promoters supports the promoters ( $\pm$  500 bp around the gene's TSS) of paused genes have a higher density of both than non-paused genes, consistent with the qualitative assessment in a previous report[145]. Comparatively, more weakly paused genes (lower median PI across all analyzed cell types for the specified organism) had a lower CpG content at promoters than more highly paused genes, suggesting that the build-up of paused Pol II may be partially a function of the promoter sequence composition in both mouse and human.

bp around TSS) GC and CpG content distribution for each pausing category against non-paused genes (see Methods). Paused genes had higher promoter GC and CpG content ( $p < 0.001$ , Mann-Whitney U test; see Figure 4-5). In a complementary analysis, we ranked all 6- and 5-mer DNA sequences by their over-representation in paused versus non-paused gene promoters (see Methods). Over-represented n-mers had both high GC and CpG content (see Figure 4-6). Whereas *Drosophila* paused promoters were enriched for more TA rich motifs including the TATA box[139, 197], AT-rich motifs were depleted in paused mammalian promoters (see Figure 4-6). Our analysis extends previous associative studies by providing a quantitative analysis of how GC and CpG promoter density relates to increasingly paused Pol II at mammalian gene promoters. These analyses also suggest that a subset of the DNA sequence elements that influence pausing differ between mammals and *Drosophila*.

### 4.3.2 Pol II Pausing Influence On Gene Expression Levels

The relationship between gene expression and pausing was addressed previously in a limited range of cell types using microarray expression data[195, 199]. To describe this relationship more comprehensively using more sensitive RNA-seq approaches, we evaluated the influence of Pol II pausing on transcription across multiple mammalian cell lines by integrating RNA-seq gene expression data with Pol II ChIP-seq. Paused genes composed the majority of expressed genes (FPKM>1) in each cell type and were a significantly higher fraction of expressed genes compared to all annotated genes ( $p < 0.001$ , proportion test; see Figure 4-7). Yet, we found no consistent expression level difference between paused

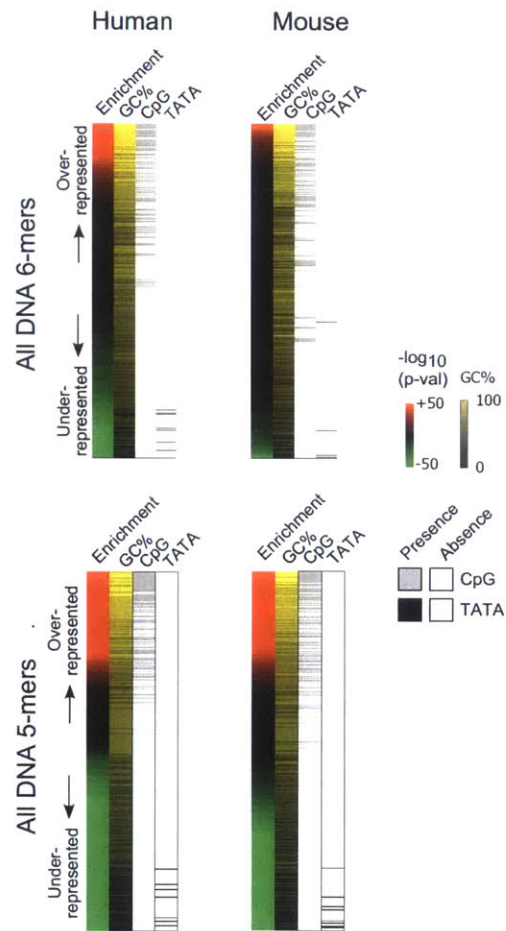


Figure 4-6: Calculating enrichment 6- and 5-mers in the promoter (+/- 500 bp around TSS) of paused genes versus non-paused suggest shows the sequence composition of genes have have significant pausing.



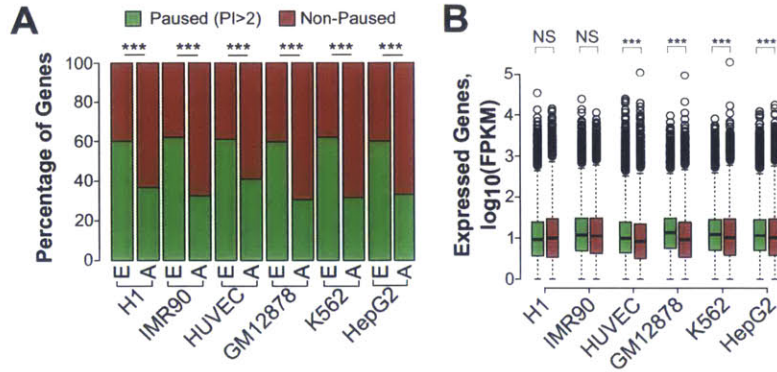


Figure 4-7: Paused and non-paused genes have a similar expression profile even while paused genes are the predominantly expressed gene in a cell type. In (A), the proportion of expressed genes (where FPKM>1) in each studied cell type tend to have a PI of at least 2. (\*\*\*)  $p < .001$ , proportion test). Yet, in (B), there is no consistent difference between the expression level of paused and non-paused genes (\*\*\*)  $p < .001$ , NS = not significant, Mann-Whitney U test), although some cancer cell lines have an upregulated set of genes with a PI>2 relative to non-paused genes.

and non-paused genes (see Figure 4-7), although particular analyzed cell types had a slight higher average expression of paused genes ( $p < 0.001$ , Mann-Whitney U test; see Figure 4-7) suggesting certain cellular states may drive increased expression.

We next modeled the trend between gene expression levels and PI, including both expressed and non-expressed genes with an assigned PI. In GM12878, PI and gene expression correlated weakly (Spearman  $r = 0.179$ ; see Figure 4-8 and Methods) but we found a surprising “hill-shaped” trend where the highest mean expression (red trend line) and the widest range of gene expression (blue density plot) occurred over intermediate PI values. Genes outside this intermediate PI range had reduced expression level. Moreover, a similar hill-shaped trend and weak correlation recurred across four of the five cell types tested (mean Spearman  $r = 0.09$ ; see Methods and Figure 4-8 and B-4), suggesting it is a common relationship at steady-state. The peak of the trend lines for IMR90, K562, and GM12878 covered a similar PI range. The peak of the H1 embryonic stem cell trend line was shifted leftward relative to the other analyzed cell types suggesting that the relationship between gene expression and pausing can change in a cell-type specific manner. In comparison to PI, gene expression was monotonically and positively correlated with TSSR, gene body, and total Pol II density (for all samples, mean Spearman  $r = 0.45, 0.5, \text{ and } 0.5$ , respectively; see Figure B-4). Together, our higher-resolution analyses indicate that Pol II density is a better predictor of steady state gene expression levels than PI. PI weakly correlated with gene expression, but very high or low PIs were linked to lower expression levels, suggesting that a balance between Pol II TSS loading and pausing release may be needed to sustain higher levels of gene expression.

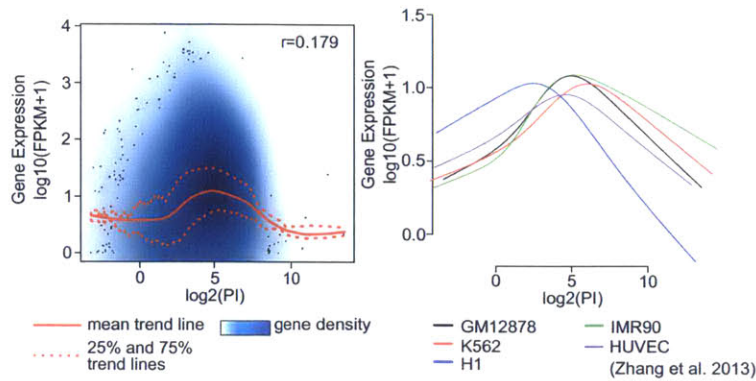


Figure 4-8: Gene expression to PI in GM12878 showed no linear trend between the two variables. However, there was a “hill-shaped” trend between the two variables, with the highest average gene expression was near the middle of the PI distribution for a cell. At the extremes, mean gene expression was lower, suggesting that too much or too little Pol II build up at TSS negatively affected gene expression. This trend was common across multiple cell types.

### 4.3.3 Pol II Pausing and Cell Population Gene Expression Variability

Pol II pausing has been recently shown to help synchronize gene expression across individual cells relative to non-paused genes in *Drosophila* embryos[193, 194, 200]. While there was not a strong relationship between gene expression and how paused Pol II is at a gene, it was explored whether pausing regulates gene expression by affecting cell-to-cell expression variation in mammals using single-cell RNA-seq data. The single-cell RNA-seq data was looked at in GM12878[201] (EBV-immortalized B-cells), and H1[202] (human embryonic stem cells) since there was matching Pol II ChIP-seq data. To test whether pausing may affect the expression variability within the cell population, the expression coefficient of variation ( $CV = \frac{\sigma}{\mu}$ ) was compared between paused and non-paused expressed genes (mean FPKM>1 across the population of individual cells in each sample). Paused genes had lower expression CV within all but the highest expression quantiles (see Figure 4-9), even as expression CV decreased with increased expression levels as previously shown[203]. There was no systematic difference in the expression levels between paused and non-paused genes in each quantile (see Figure 4-10), suggesting that expression level differences did not cause the observed difference in CV between paused and non-paused genes. This analysis suggests that paused Pol II stabilizes gene expression level across individual cells in a population.

### 4.3.4 High Pol II TSS density promotes pausing release

A novel strategy was developed to visualize the interaction between Pol II TSSR density, gene body density, PI, and an additional variable such as gene expression (see Figure 4-11). Each gene was plotted by its Pol II TSSR and gene body density and colored by gene expression. In these plots, the PI is given by the position of a gene along the left-to-right



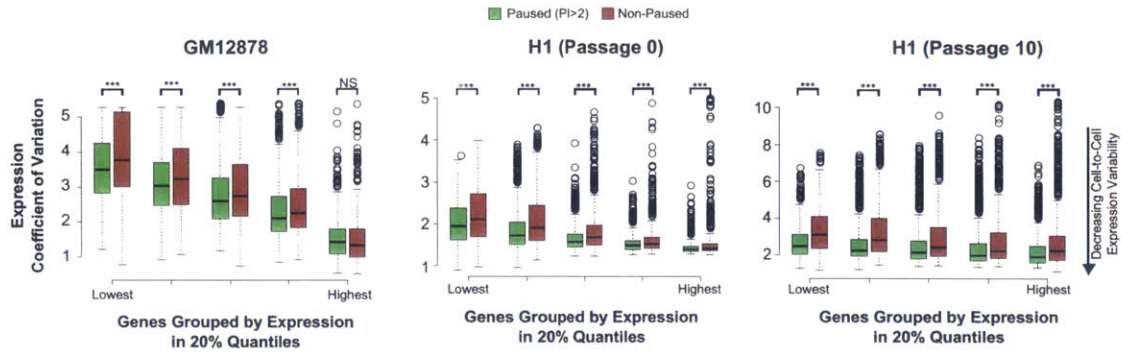


Figure 4-9: Paused genes have a lower cell-to-cell expression variability than non-paused genes for the same population mean expression level across cell types. (Mann-Whitney U test, NS = not significant, \*\*\*  $p < .001$ )

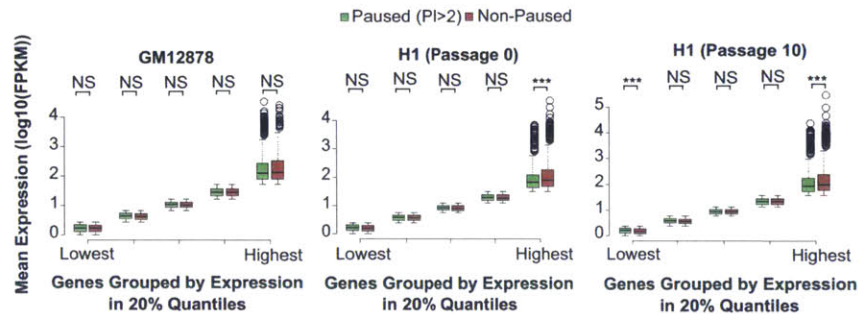


Figure 4-10: With the expression quantiles in each single cell experiment, there was no systematic different in the distribution of mean, population wide expression levels for paused and non-paused genes in the same quantile. (Mann-Whitney U test, NS = not significant, \*\*\*  $p < .001$ )

diagonal. Modeling the trend between TSSR and gene body Pol II density uncovered a novel biphasic relationship (red line, see Figure 4-11) in which the trend line turned upwards and became steeper at an “inflection point” (pink arrow). The trend was observed in all 6 samples evaluated, and in each case the inflection point occurring at high, but slightly different, TSSR Pol II densities (see Figure 4-12). The inflection point indicates that increasing Pol II TSSR loading above a threshold level causes an overall greater increase in pausing release relative to genes of lower TSSR Pol II densities.

Functionally, genes past the trend line inflection point within each cell type shared similar expression and GO term profiles with each other. These genes were over-represented for highly expressed genes (FPKM>1000) compared to all expressed genes within a cell type ( $p < 0.001$ ; proportion test; see Figure 4-12). These genes were enriched for diverse functional terms across cell types, including “cellular metabolic process”, “biosynthetic

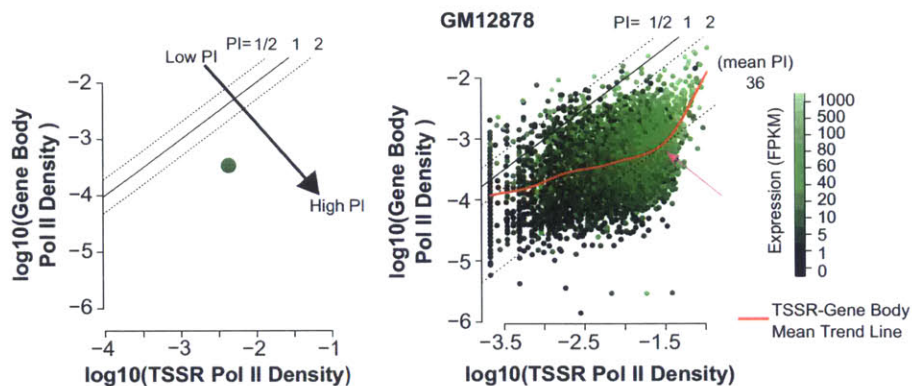


Figure 4-11: To analyze the nature of Pol II activity and the PI, a visualization technique was developed to plot each gene by its Pol II TSSR and gene body density. Then the PI can be visualized in the left to right diagonal. Each point/gene can be colored relative to a third variable, such as gene expression. When plotting all genes in GM12878 as such

process”, and “gene expression” (see Figure 4-13). These results suggest that elevated pausing release for genes to the right of the inflection point contributes to expression of many highly expressed genes, which often participate in housekeeping activities.

An increase in machinery related to Pol II pausing and pausing release may drive elevated pause release with increased Pol II TSSR density to the right of the inflection point, so the promoter occupancy of CCNT2 and CDK9, components of P-TEFb, and NELF were analyzed for its relationship to PI. All three components strongly correlated with TSSR density against the Pol II in the corresponding cell type ( $r = .61, .61, .65$ , respectively, Spearman correlation), where our result for NELF was consistent with previous reports[195, 197], where comparatively each component had a weaker correlation with PI ( $r = .26, .20, .29$ , respectively). The analysis was confirmed by visualizing each factor’s occupancy of gene promoter as a function of the gene’s TSSR and gene body Pol II density (see Figure ??). These analyses revealed that promoter occupancy of all three factors, especially P-TEFb, strongly correlated with increased Pol II TSSR density, varied little with Pol II gene body density, and were all strongest past the inflection point. This also suggests that P-TEFb is well positioned to elevate pause release for the genes to the right of the inflection point.

To mechanistically assess which factors were responsible for establishing the biphasic relationship, datasets where Pol II ChIP-seq was performed in cells treated with inhibitors of pausing-related machinery were analyzed. First, it was tested whether P-TEFb is required for elevated pause release to the right of the inflection point. Treatment of both mouse embryonic stem cells (mES) [195] and human lung fibroblasts (IMR90)[204] with flavopiridol (FP), an inhibitor of P-TEFb activity[140], reduced the steeper portion of the

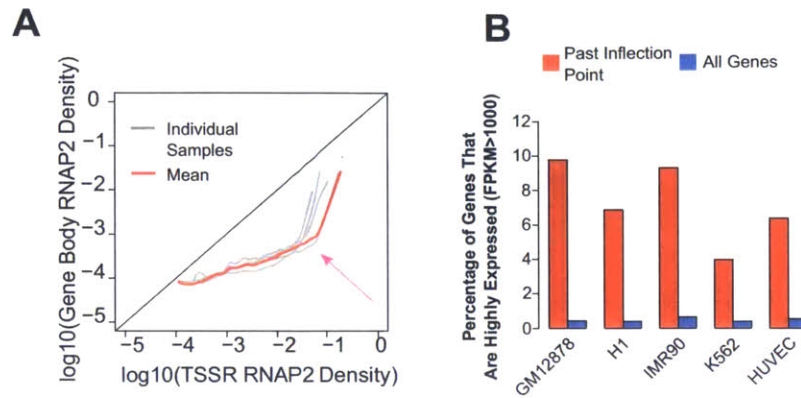


Figure 4-12: The observed inflection point in the Pol II TSSR-gene body trend curve re-occurred across multiple cell types and had an enriched amount of highly expressed genes (FPKM>1000).

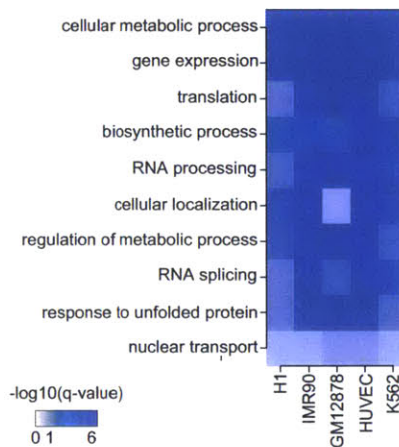


Figure 4-13: Recurrently enriched GO biological process terms for genes past the inflection point in their respective cell type. Generally, most GO biological process terms observed recurrently were for general functions of the cell (*e.g.*, gene expression).



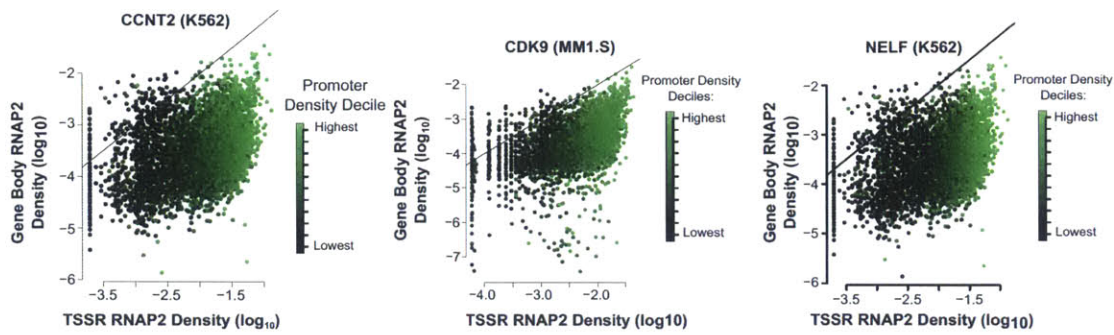


Figure 4-14: Compared to the deposition of H2A.Z, the multiple pausing related components only varied in deposition with changes in Pol II TSS density (and likely initiation), but changes in how much Pol II entered the gene body (as a proxy for how much Pol II is elongating) did not have a strong effect.

trend line to the right of the inflection point (see Figure 4-15). This suggested that P-TEFb activity is necessary for high TSSR Pol II density to stimulate Pol II accumulation on the gene body. FP treatment of IMR90 also increased the maximal TSSR Pol II density (see Figures 4-15 and 4-16), suggesting that elevated Pol II pause release to the right of the inflection point in untreated cells was not due to TSSR saturation with Pol II. Secondly, the effect of antagonizing pausing release factors NELF-A[195], BRD4[117], and ELL3[205] with siRNA or small molecules was also examined. These treatments had subtle effects on the trend curve but did not reduce the inflection point (see Figure 4-15). This suggested that these other factors are not essential determinants of the biphasic TSSR to gene body Pol II density relationship.

These analyses highlight a novel relationship between initiation and pause release, in which gene body Pol II density rapidly increases at high TSSR Pol II densities. This may be driven by elevated P-TEFb density, which increases with higher TSSR Pol II density, but is not specific with any particular studied pausing release factors known to drive increased elongation through increasing P-TEFb activity[117, 205].

#### 4.3.5 Many stimulus-responsive genes are paused and have lower PI prior to stimulation

Pol II pausing was initially hypothesized to regulate stimulus-responsive gene expression change [77, 132, 206, 207], but a recent study challenged this idea and suggested that responsive genes tended to lack paused Pol II prior to stimulation[142]. Accordingly, Pol II pausing's role in stimulus-induced alterations of gene expression was further investigated here. We analyzed the time-course of HUVEC response to stimulation with vascular endothelial growth factor A (VEGFA)[150]. Based on gene expression fold-change, gene expression changes were grouped into early upregulated, late upregulated, downregulated, and non-responsive gene sets (see Figure 4-17 and Methods). In contrast to previous reports[142, 207], the majority of genes in each set were paused prior to stimulation (see

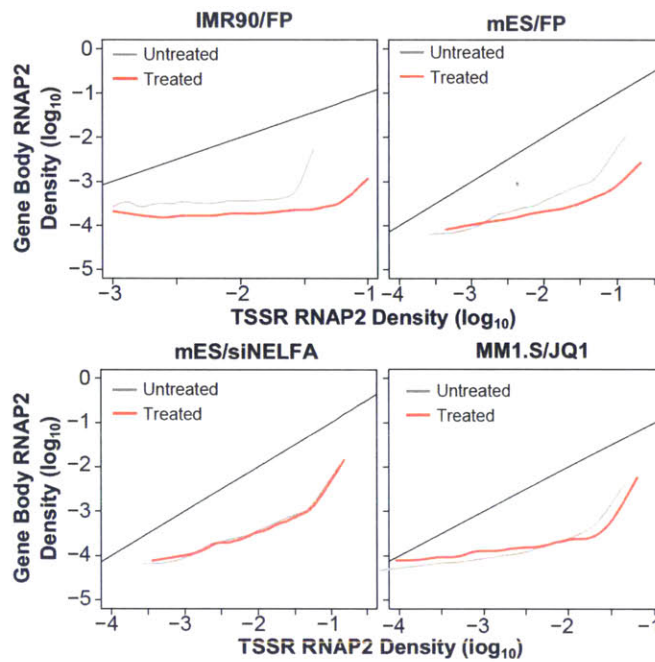


Figure 4-15: When analyzing Pol II samples with FP treatment, the inflection point is dramatically reduced in both samples, confirming that the inflection point is driven by increased P-TEFb activity. However, none of the other treatments significantly reduced the inflection point, suggesting that the inflection point may not be driven by a particular driver of pausing release.

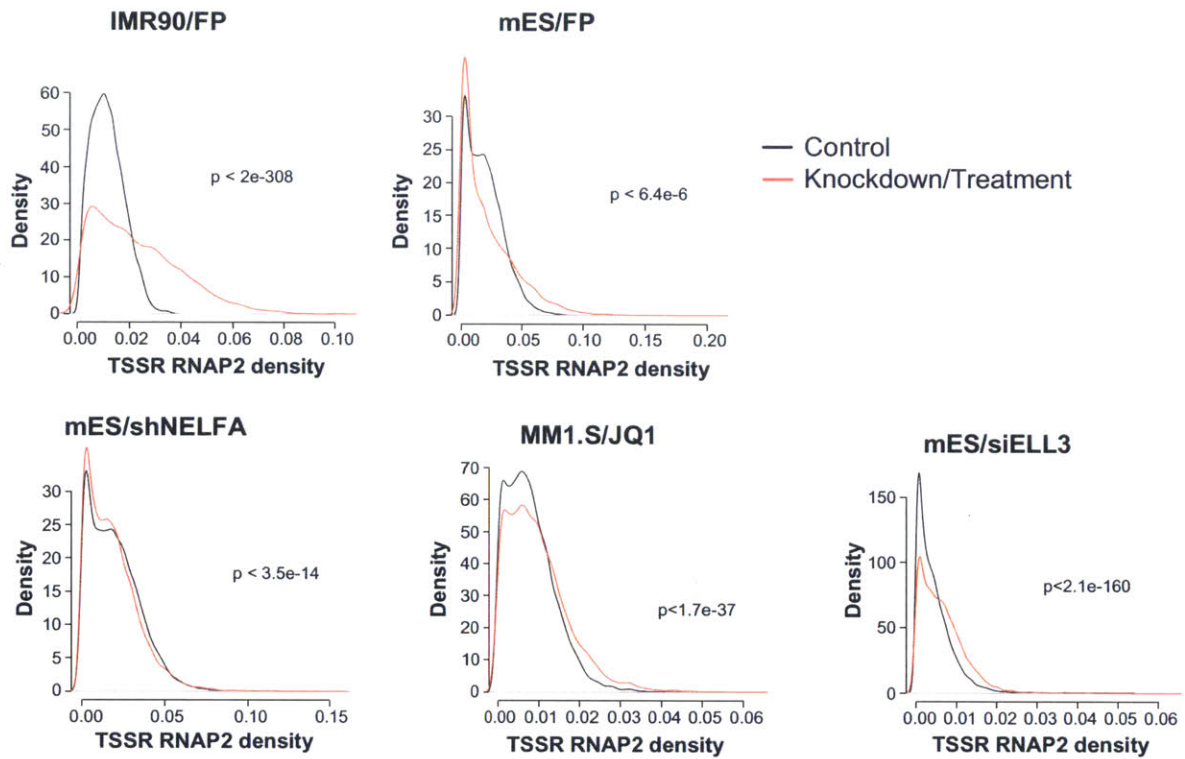


Figure 4-16: Inhibition of pausing release through various inhibitions and comparing change in Pol II TSS density shows that the Pol II TSS density can increase.



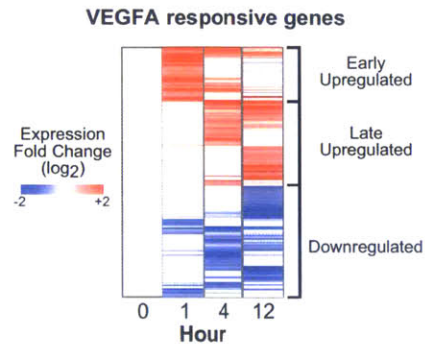


Figure 4-17: VEGFA responsive gene sets based on fold-change in gene expression. Only genes with a minimum 2-fold gene expression change (up or down) are included here.

Figure 4-18). Surprisingly, genes responsive to VEGFA had a lower had a lower average PI than non-responsive genes ( $p < 0.001$ , Mann-Whitney U test; see Figure 4-19). To determine if this observation extended to other stimuli and cell types, two additional analyses were performed. First, genes in HUVECs that responded to IL-4 stimulation were identified from time-series microarray expression data[208] (see Methods). These IL-4 responsive genes also had a lower average PI than IL-4 non-responsive genes in unstimulated HUVECs based on our baseline (Hour 0) HUVEC Pol II ChIP-seq data ( $p < 0.001$ , Mann-Whitney U test; see Figure 4-19), suggesting the effect is likely independent of the underlying stimulus. Second, IMR90 fibroblasts stimulated for 1 hour with tumor necrosis factor alpha (TNFa)[204] were examined. Prior to stimulation, TNFa responsive genes again had a lower average PI than non-responsive genes ( $p < 0.001$ , Mann-Whitney U test; see Figure 4-19). This was also the case for TNFa “primary response genes”, defined as those genes that are differentially expressed without the need for new protein synthesis ( $p < 0.001$ , Mann-Whitney U test; see Figure 4-19). Together, these analyses suggest that many stimulus-responsive genes are paused but have lower than average PI pre-stimulation.

### 4.3.6 Pausing release selectively regulates rapid, signal-induced gene expression change

Since most responsive and non-responsive genes were paused prior to stimulation (see Figure 4-18), it was then asked whether altered Pol II pausing release regulates stimulus-induced expression changes. In VEGFA-stimulated HUVECs[150], early upregulated genes shifted leftwards to become less paused at one hour ( $p < 0.001$ , Mann-Whitney U test; see Figure 4-20) and then shifted rightward back towards their basal, more paused state at 4 and 12 hours. This time course for PI change paralleled the time course of gene activation of these early upregulated genes (see Figure 4-17). Neither the late up-regulated, down-regulated, nor non-responsive genes had a similar significant PI distribution shift during the time course (see Figure 4-20).

The requirement for pausing release in driving responsive gene change in VEGFA-

**VEGFA expressed genes RNAP2 bound and paused before stimulation (Hour 0)**

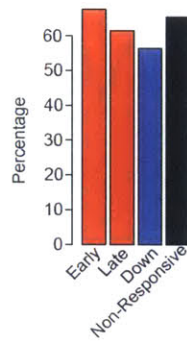


Figure 4-18: Fraction of expressed, VEGFA-responsive and non-responsive genes that are paused ( $PI > 2$ ) at Hour 0. Unlike a previous report[207], the build-up of paused Pol II at TSS does not define different types of responsive genes or even between responsive and non-responsive genes.

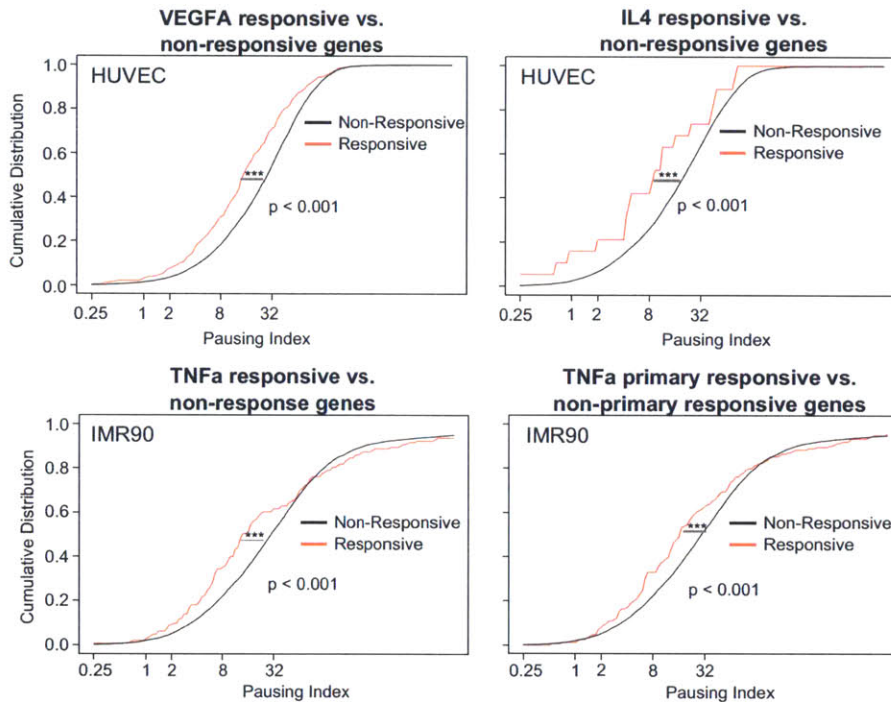


Figure 4-19: Responsive genes had a lower average PI compared to non-responsive genes within the same cell type and stimulus. Mann-Whitney U test, \*\*\*  $p < .001$ .



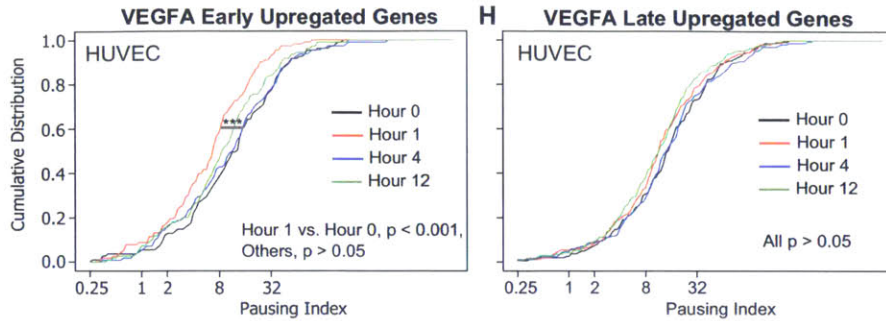


Figure 4-20: Early upregulated genes (as defined in Figure 4-17) are differentially released from pausing in response to VEGFA compared to late upregulated genes, which on average have a similar PI distribution throughout the time course. The early upregulated genes also are only released for a short time period before returning towards baseline.

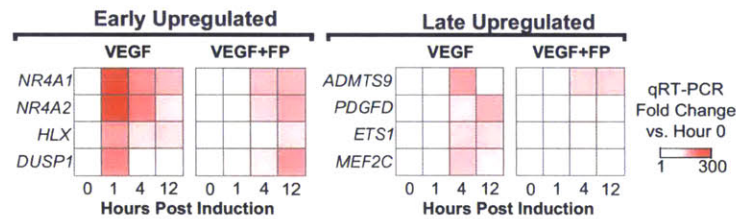


Figure 4-21: Expression measurement by RT-PCR of select VEGFA responsive genes in HUVECs with and without FP treatment. The FP treatment dramatically inhibited VEGFA responsive genes. Interestingly several of these VEGFA responsive genes begin to upregulate later in the time course. (qRT-PCR performed by Dr. Bing Zhang.)

stimulated HUVECs further was mechanistically tested for by measuring the effect of FP on VEGFA activation of selected early and late responsive genes. FP treatment dramatically suppressed the strong VEGFA-induced upregulation of early genes at 1 hour (see Figure 4-21), while the “tail” of their activation at 4 and 12 hours was less strongly affected. FP treatment also affected activation of late genes at 4 and 12 hours (see Figure 4-21), although this effect was less pronounced than the effect on rapid induction of early genes. While many responsive and non-responsive genes are paused, this analysis suggests that increased pausing release conditionally regulates rapid, signal-induced gene activation.

### 4.3.7 Pol II pausing relationship to the local chromatin landscape

First, exploration for recurrent relationships between nucleosome and histone modification occupancy with respect to a gene’s PI was performed. In *Drosophila*, paused genes tended to have a greater nucleosome depletion at TSS[197]. In GM12878 and K562, monococcal-nuclease digested chromatin followed by high-throughput sequencing (MNase-seq) from ENCODE[165] was used to map nucleosome density around each TSS with Pol II signal.

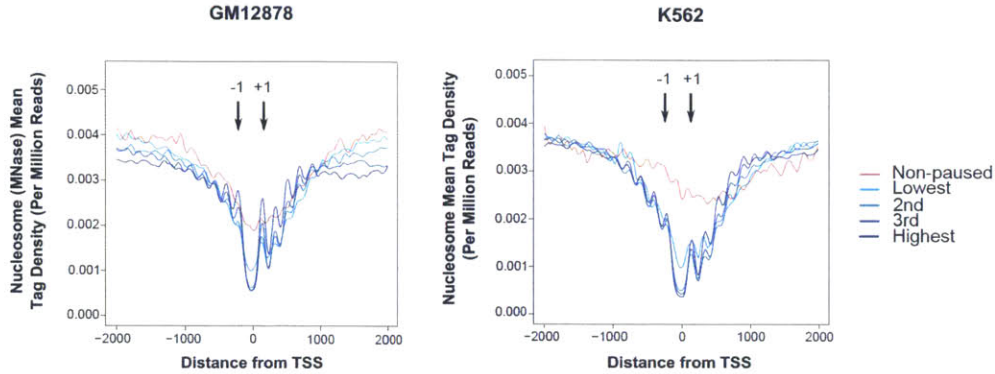


Figure 4-22: Paused genes have a greater nucleosome depletion at TSS of expressed genes in GM12878 and K562. Nucleosome density was measured by MNase-seq genome-wide and plotting the density on average. Although there was a major difference between paused and non-paused expressed genes, there was little difference in the nucleosome depletion at TSS with varying levels of the PI.

Consistent with previous observations in *Drosophila*, paused human promoters had a significant nucleosome depletion at TSS compared to non-paused genes (see Figure 4-22). However, between different PI strengths of paused genes there was not a significant difference in nucleosome depletion. This may suggest that the establishment of Pol II even a weakly paused state may be sufficient to establish the nucleosome depletion, consistent with NELF knockdown that do not cause significant shifts in nucleosome depletion at paused genes in *Drosophila*[147].

Paused genes tended to have a strong +1 and -1 nucleosome around TSS (see Figure 4-22), suggesting that these nucleosomes may have a regulatory relationship with pausing and may be differentially linked with particular histone modifications based on how paused Pol II is. Using previously published histone modification and H2A.Z data from ENCODE across multiple cell types[50] against matched Pol II ChIP-seq data, we identified positive correlations between PI and fold enrichment for several histone marks associated with active genes that were maintained across samples (see Figure 4-23 and Methods). This analysis revealed that H2A.Z, a histone variant essential for lineage commitment and embryonic development[42], positively correlated with PI across the samples analyzed. This was surprising because in *Drosophila* H2A.Z at TSS destabilized promoter nucleosomes and consequently was anti-correlated with Pol II pausing[41, 209]. H2A.Z was more closely correlated to PI than to either TSSR or gene body Pol II density (for K562  $r = .55$ ; see Figure 4-24). In contrast, the TSSR density of histone H3.3[210], another histone variant, in HeLa was not correlated with PI ( $r = .13$ ; see Figure 4-24). Furthermore, H2A.Z occupancy of the -1 and +1 nucleosome positions positively correlated with increasing PI (see Figure 4-25). Together, these data identify a positive correlation between H2A.Z and Pol II pausing in mammalian cells.

It was tested if H2A.Z function to destabilize promoter nucleosomes and antagonize pausing, previously reported in *Drosophila*[41], was conserved in mammals. In murine ES



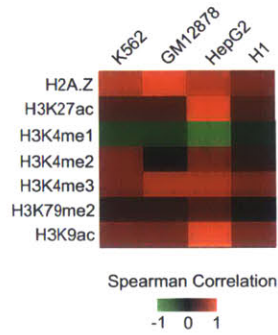


Figure 4-23: Estimated Spearman correlations for multiple chromatin markers TSS density against the gene's PI across multiple cell types.

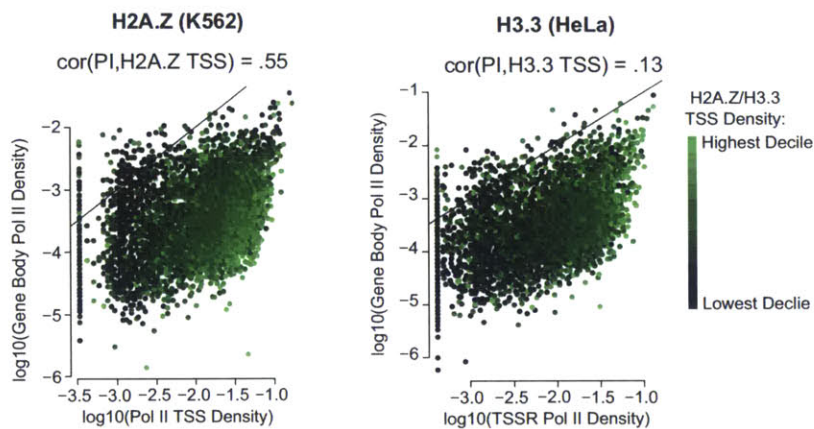


Figure 4-24: Using the visualization developed in Figure 4-11, plotting H2A.Z TSS density in K562 with respect to Pol II TSSR and gene body density visually shows that H2A.Z TSS density strongly increases in only the direction of the increasing PI (along the left to right diagonal) rather than with Pol II TSSR or gene body density itself. Comparatively, analyzing H3.3 TSS density in HeLa shows that the deposition of H3.3 is different that H2A.Z, suggesting that nucleosome turnover itself does not cause the observed effect in H2A.Z.

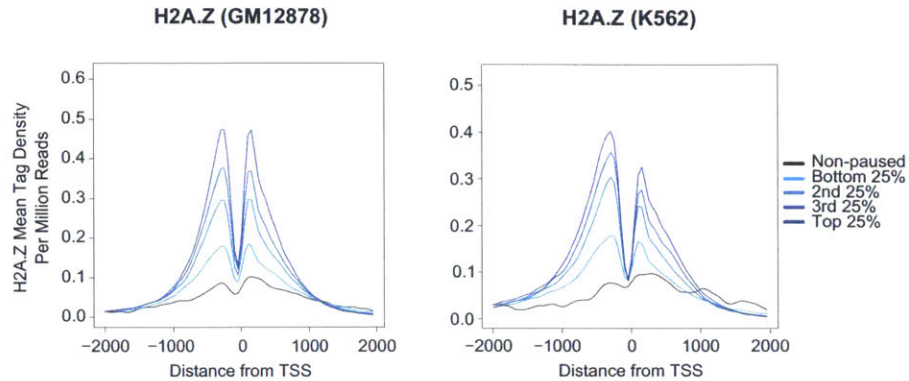


Figure 4-25: For GM12878 and K562 H2A.Z data, the H2A.Z deposition around TSS increases with increasing PI, consistent with the observations in Figures 4-23 and 4-24).

cells, H2A.Z knockdown globally increased promoter nucleosome density, and moreover the increase in promoter nucleosome density was the greatest at the most paused genes (see Figure 4-26). To further interrogate a potential functional relationship between H2A.Z and Pol II pausing, H2A.Z was knocked down with siRNA in MCF7 breast cancer cells (see Figure B-6 and Methods) and then measured changes in pausing by performing Pol II ChIP-seq. H2A.Z knockdown increased Pol II pausing ( $p < 0.001$ , Mann-Whitney  $U$  test; see Figure 4-27), consistent with the effect of H2A.Z depletion on pausing in *Drosophila*[41]. For genes with increased pausing upon H2A.Z depletion in MCF7, we analyzed the change in Pol II density at TSSR and gene body. Among most genes with greater than two-fold change in either category, Pol II density either increased in the TSSR or diminished in the gene body, but not both (see Figure 4-28). Interestingly, the subset that decreased gene body Pol II density upon H2A.Z depletion was distinguished by a significantly lower baseline PI compared to all genes or the subset with increased TSSR Pol II density (see Figure 4-28). These data suggest that there are likely at least two distinct mechanisms by which H2A.Z modulates Pol II pausing. Furthermore, our analyses indicate that H2A.Z destabilizes nucleosomes and antagonizes Pol II pausing in both *Drosophila* and mammals.

We asked how H2A.Z antagonized pausing yet its promoter occupancy positively correlated with Pol II pausing. We hypothesized that increased pausing stimulates H2A.Z promoter deposition, resulting in a negative feedback loop to control pausing. To test this hypothesis, MCF7 cells were treated with FP to block Pol II pause release. Then, H2A.Z enrichment was measured at promoters by ChIP-qPCR. Consistent with our hypothesis, we observed that FP treatment increased H2A.Z enrichment overall at the test promoters (all promoters together comparing treated vs. un-treated,  $p < .001$ , t-test; see Figure 4-29). Together the data suggests that H2A.Z destabilizes nucleosomes and antagonizes pausing in mammals, consistent with previous reports[41, 42, 209]. Unlike in *Drosophila*, H2A.Z correlates with Pol II pausing in mammals, potentially due to a negative feedback loop in which pausing increases H2A.Z promoter deposition.



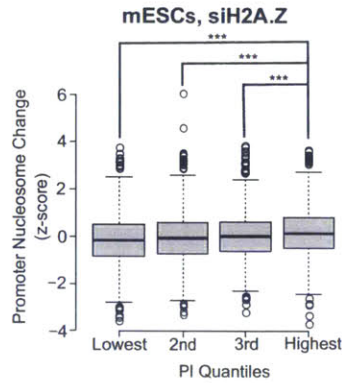


Figure 4-26: The nucleosome density around TSS for paused genes in mES increases highest at the most paused genes. ( $*** p < .001$ , Mann-Whitney U test)

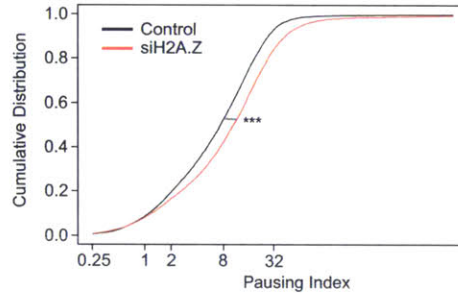


Figure 4-27: When MCF7 cells are transfected with siRNA against H2A.Z, the PI distribution shifts rightward (average PI increases, as a result) in response to the H2A.Z knock-down. ( $*** p < .001$ , Mann-Whitney U test)

**Changes in RNAP2 density at MCF7 genes with increased PI upon H2A.Z depletion**

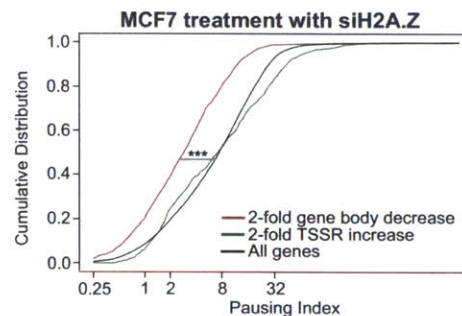
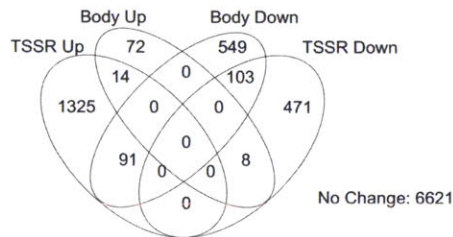


Figure 4-28: When H2A.Z was knocked down by siRNA in MCF7 cells, the change in Pol II did not occur uniformly across all genes. Predominantly, genes either lost Pol II density in its gene body or doubled the Pol II density at the TSS. Additionally, the determination of how Pol II changed at a gene appeared to be related to its pre-H2A.Z knockdown PI.

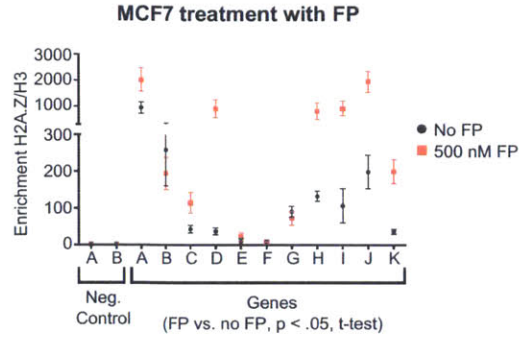


Figure 4-29: Comparison of H2A.Z enrichment relative to H3 density at selected promoters using ChIP-qPCR. After treating MCF7 cells with FP, 8 out of 11 sites had significantly increased H2A.Z deposition around the target gene's TSS. Grouping all the non-control sites together, these regions had a significantly higher fraction of H2A.Z enrichment relative to H3 ( $p < .001$ , t-test). It appeared the inhibiting Pol II pausing release allowed for an increase in H2A.Z buildup at the TSS. (ChIP-qPCR result provided by Dr. Bing Zhang.)

### 4.3.8 Pol II pausing's relationship to chromatin topology

It has been recently shown that distal regulatory elements can regulate Pol II pausing release[117, 205], in addition to Pol II initiation. Additionally, distal regulatory elements can target multiple genes through long-range interaction as well as enhancers and promoters can interact with multiple other enhancers and promoters, respectively[211]. This raises the question whether genes linked to the same regulatory regions by long-range interactions may co-regulation of Pol II pausing several genes. First, the genes within topological associating domains (TADs) as identified in Hi-C interaction data[122] were analyzed. Chromatin within TADs are highly interactive, suggesting that genes and enhancers within a TAD are often in close physical proximity[122]. To test whether genes within a TAD had a similar PI, which suggests that their pausing state is co-regulated, the PI coefficient of variation (CV) was estimated for each TAD and then calculated the mean PI CV across all TADs. This observed mean PI CV across TADs was compared to a permuted background controlling for gene expression and gene number per TAD (see Methods). The observed mean PI CV was significantly lower than expected (see Figure 4-30), suggesting that genes within a TAD tended to have more similar PIs than across TADs. A complementary analysis was performed using Pol II ChIA-PET data to define enhancer-promoter interactions (see Methods). Using a similar procedure to analyze the TADs, genes interacting with the same enhancer had a significantly lower mean PI CV than expected based on the permuted background (see Figure 4-31). These analyses suggest that distal regulatory elements also have a coordinating effect on regulating Pol II pausing across multiple genes.



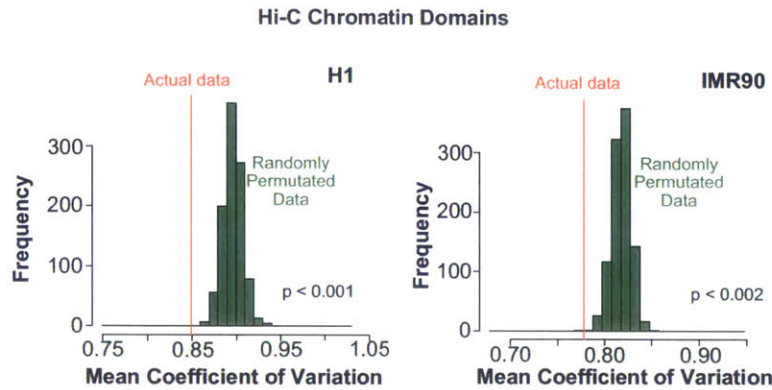


Figure 4-30: Across H1 and IMR90 cell lines, the mean PI coefficient of variation for genes within the same TAD had was significantly lower than the randomly generated background.

## 4.4 Discussion

This approach of analyzing multiple, previously unconnected datasets from diverse mammalian cells and tissues generated novel insights and new directions about the function of Pol II pausing. While many studies have elucidated mechanisms of pausing and pausing release[77, 117, 136, 139, 143, 145, 193, 194], this was a more genomic approach to understand the breadth and possible functional and regulatory roles for Pol II pausing in mammals. This study provides new direction on the establishment of Pol II pausing in mammals while highlighting potential functional roles in gene regulation.

This analysis illuminated the prevalence of Pol II pausing in mammals. Previous studies estimated the number[77, 143–145] and functional enrichments of paused genes[142–144], but only in a few cell types. This broad analysis across many cell types and found it to be consistent across samples, even in cancer cell lines. Interestingly, some enriched functional categories differed between species. For example, developmental genes were less paused in mammals than in *Drosophila*[144]. Our study confirms the widespread nature of Pol II pausing and suggests that its specificity changed over evolutionary time.

The span of pausing across cell types suggested that a gene's primary sequence helps to determine the strength of Pol II pausing. Previous studies uncovered multiple primary sequence features that are enriched at paused promoters in *Drosophila*, such as the TATA box[139, 197] and high promoter GC and CpG content[198]. In humans, annotated CpG islands strongly associated with paused promoters (Core et al. 2008), but it was unknown what other features related with paused promoters in *Drosophila* are conserved in humans. We found that both increased promoter GC and CpG content associated with increased pausing across human and mouse samples. High promoter CpG and GC content likely increases Pol II initiation by depleting nucleosomes near the TSS, without affecting downstream (+1) nucleosome density at the shore of the high CpG/GC region[212, 213]. However, enrichment of the TATA motif in paused *Drosophila* promoters[139, 197] did not

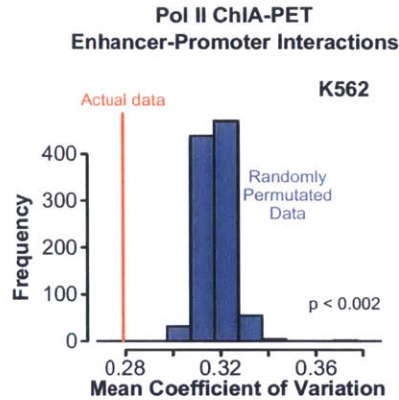


Figure 4-31: Similar to the analysis in Figure 4-30, the mean PI coefficient of variation for genes associated with the same enhancer defined through Pol II ChIA-PET long-range interactions was also significantly lower than the permuted background ( $p < 0.012$ ).

carry over to mammals. This finding highlights species-specific adaptations to regulate Pol II pausing. Furthermore, our study opens new avenues towards understanding the engineering and evolution of paused promoters.

There was a complex relationship between Pol II pausing and gene expression. An initial function ascribed to Pol II pausing was “pre-loading” of Pol II on promoters of rapidly induced genes to enable their rapid stimulus-triggered expression[77, 132], but this has been called into question[77, 142]. While some studies have suggested that Pol II pausing inhibits expression (Schones et al. 2008), others have found that paused Pol II is productive[77, 143, 214]. Our analyses supported the productive potential of paused genes, and the hill-shaped PI-gene expression pattern indicated that a wide range of PIs is compatible with a range of expression. Only when a gene reached relatively high or low PIs within a cell type did we observe reduced gene expression, suggesting that adequate Pol II promoter loading and pausing release are required to sustain robust gene expression. Across individual cells, there was lower cell-to-cell variability in the expression of paused versus non-paused genes, suggesting that pausing may help dampen transcriptional bursting[194, 215]. Both responsive and non-responsive genes tended to be paused prior to stimulation, but the baseline PI distribution of responsive genes was shifted towards lower PI values compared to non-responsive genes. Stimuli triggered rapidly induced genes to transiently reduce their pausing further, implicating stimulus-induced pause release in the rapid transcription of this class of genes. Stimulus-induced pause release was not observed for less rapidly responding groups of genes, suggesting selectivity of this mechanism for rapid transcriptional responses.

There was novel interplay between Pol II pausing and chromatin structure. Whereas H2A.Z occupancy anti-correlated with Pol II pausing in *Drosophila*[41, 209], in mammals we found that H2A.Z promoter occupancy positively correlated with Pol II pausing. These analyses suggest that the fundamental mechanism of H2A.Z action on pausing “reduction of the nucleosomal barrier to productive elongation by decreasing nucleosome stabil-



ity[41–43] is likely conserved between mammals and *Drosophila*. Additional layers of regulation must be present in mammals to reconcile this mechanism with the positive correlation that we observed between H2A.Z and pausing. One hypothesis, supported by our finding that FP increased H2A.Z density, is that Pol II pausing promotes H2A.Z deposition. This would create a negative feedback loop in which Pol II pausing enhances H2A.Z deposition, which then limits Pol II pausing. Additional studies are needed to further test this hypothesis and to account for the opposite correlations between H2A.Z and pausing in *Drosophila* compared to mammals. In a more global context, the Hi-C and ChIA-PET analysis suggests that the topological structure of chromatin may also regulate Pol II pausing by coordinating the strength of pausing across multiple genes. Further investigation is required to delineate how chromatin topology regulates Pol II pausing and pause release.

## 4.5 Additional Methods

### 4.5.1 ChIP-seq and RNA-seq analysis

All ChIP-seq data was aligned to human reference genome hg19 or mouse reference genome mm9 as appropriate using bowtie2[60].

All RNA-seq data was aligned to the RefSeq human or mouse genomes as appropriate using tophat2[189]. FPKM quantification was performed using cufflinks.

### 4.5.2 Calculating Pol II Pausing

Pol II pausing was estimated by calculating the Pausing Index (PI), based on recent papers that have used a similar quantity to study the nature of Pol II at the gene. The PI was calculated by first estimating the TSSR Pol II density (-50 to +300 bp around the gene's TSS) and the gene body Pol II density (+300 past TSS to +3 kb after the TES). These density estimates are normalized the input read density (by subtracting it), ChIP library size, and then to the size of the region. Then the PI is the ratio between the two densities as follows for a gene  $g$ :

$$PI_g = \frac{TSSR_g}{Body_g}$$

A graphical summary of the calculation for the PI is in Figure 4-1. This PI index measures the relative distribution of Pol II throughout the body, where a higher PI suggests Pol II sits at TSS in paused state more often. The RefSeq gene annotations for humans and mouse were used to define the regions described above. If a gene had multiple annotated TSSes or multiple different TESes, then the TSS with the highest H3K4me3 and/or Pol II was chosen and the farthest TES for isoforms relating to that TSS was taken.

### 4.5.3 GO Analysis for paused genes

Took the top and bottom 25% of paused genes per cell type. Calculated the GO enrichment based on the HGNC name using the *topGO* R package[190]. The p-values for each GO biological process term were calculated using the hypergeometric test. Resulting p-values

were corrected for multiple testing by converting them into q-values with the *qvalue* R package[191].

#### **4.5.4 Estimated mean PI coefficient of variation for genes within Hi-C TADs and ChIA-PET interactions**

Topological associating domains (TADs) were obtained through Dixon *et al.*[122] for H1 and IMR90. Since the coordinates for the TADs were mapped to the hg18 reference genome, the BED files containing the coordinates were re-mapped to the hg19 reference genome using *liftOver*. Peak calls and long-range interactions for the Pol II ChIA-PET data for K562 were obtained from Li *et al.*[211].

For the TADs, using matching Pol II ChIP-seq samples from ENCODE, all genes were grouped that had an assigned PI within a TAD. Although there was not a strong correlation between gene expression level and PI, the weak correlation with gene expression was corrected for by fitting a smooth spline between the two variables and subtracting the predicted PI based on the gene expression level from the estimated PIs such that the two variables became uncorrelated. I removed TADs that had less than 3 genes in them from further analysis. Within the observed data, I calculated the PI coefficient of variation (CV) for each TAD and estimated the mean PI CV for all TADs. In order to generate a background distribution to compare this value against, I shuffled genes across the TADs at random but preserving the number of genes per TAD. I then calculated the PI CV as described for the actual data, taking the mean PI CV for the background distribution. This was repeated over 1,000 permutations.

For the Pol II ChIA-PET data in K562, the PI CV p-value was calculated similarly. The key difference was that instead of grouping genes by TAD, they were grouped by having a long-range interaction with an enhancer as defined by the Pol II ChIA-PET peaks. Enhancers were defined as at least one end of the ChIA-PET loop not falling within an annotated RefSeq promoter. After grouping the genes into these sets, the calculation followed similarly to the above description using TADs.

# Chapter 5

## Conclusions and Future Directions

### 5.1 Dynamic Chromatin Changes at Signal-Responsive Regulatory Elements

This study revealed that dynamic H3K27ac remodeling is an important step in regulating the temporal activity of VEGFA-responsive regulatory elements. H3K27ac was surprisingly dynamic over the time course such that it had multiple temporal patterns in response to VEGFA, suggesting that the chromatin landscape can rapidly remodel over short periods of time. While the timing of H3K27ac change correlated with nearby temporal changes in gene expression and chromatin looping, it remains unclear whether H3K27ac remodeling is necessary for changes in VEGFA-responsive regulatory element activity or a secondary effect. Evidence to support its critical function came from blocking the acetyltransferase of p300, which inhibited dynamic H3K27ac and changes in gene expression and chromatin looping. However, p300 can acetylate other proteins aside from the N-terminal tail of histone H3, so inhibition of p300's acetyltransferase may inhibit VEGFA-responsive acetylation of other, non-histone proteins. It is possible that the acetylation of a non-histone protein by p300 is necessary to drive angiogenesis and changes in H3K27ac are secondary. Hence, further experiments are needed to test the necessity of dynamic H3K27ac at VEGFA-responsive regulatory elements. This would require separating the function of p300 to acetylate non-histone proteins and histone H3. This could be accomplished using targeted CRISPR-based techniques that would recruit histone deacetylases[216] to VEGFA-responsive regulatory elements to counteract only the histone acetylation of p300 to see if this inhibited angiogenesis, which would suggest that H3K27ac remodeling is necessary. It is unlikely that H3K27ac remodeling is sufficient to cause VEGFA-responsive gene expression change without accompanying transcription factor binding change. But it seems likely that H3K27ac remodeling is necessary to temporally modify chromatin accessibility to facilitate changes in transcription factor and other protein binding to those regions.

A novel temporal H3K27ac signature for late-activating regulatory elements was also discovered in this study. In the H4-12 cluster, H3K27ac was initially depleted at baseline before VEGFA-responsive H3K27ac remodeling filled it in after Hour 4. These sites were functionally important in VEGFA-response because of their correlation with late-activating

changes in gene expression and chromatin looping. Yet, this did not appear to be caused by changes in nucleosome positioning, since the enrichment of H3 and H3K4me2 at these sites was stable over time. Hence, any chromatin remodeling at these sites appeared specific to H3K27ac, and possibly other histone acetylations on H3 caused by p300[162]. This temporal chromatin signature may identify similar regulatory elements in other cell types, which could be tested by analyzing publically-available H3K27ac data sets (*e.g.*, ENCODE[50]) for depletions of H3K27ac similar to H4-12 sites. However, in order to find this depletion, it may be necessary to grow cells in a similar manner[150]. Overall, this would suggest that H4-12 similar sites are a class of regulatory elements whose purpose may lie in driving delayed activation of gene expression. This would then suggest that further studies would be needed to compare it to the function of latent enhancers[85]. Because both of these types of regulatory elements have delayed activation in response to an external stimulus, it would be necessary to understand the differences between these two potentially different types of regulatory elements. In particular, latent enhancers occurred at H3K27ac-absent regions whereas H4-12 sites often had other H3K27ac enrichment within the vicinity. This suggests different mechanisms of activation between these two types, and that some process may specifically inhibit H3K27ac deposition at H4-12 sites rather than lacking enrichment.

By analyzing H3K27ac dynamics, many candidate angiogenesis-related regulatory elements were identified. This also suggests a new strategy for identifying both the enhancers active in a signal-response pathway and their temporal activity. Changes in H3K27ac have been used to identify signal-responsive enhancers, but often these studies focused on two time points (*i.e.*, case versus control)[85, 125, 131]. Yet, this study shows that H3K27ac rapidly remodeled in multiple temporal patterns over a 12 hour time course, suggesting that comparing changes in H3K27ac between two time points may miss a significant number of important regulatory elements in the pathway of interest. Hence, it seems important to consider analyzing dynamic changes in H3K27ac over a longer time course in order to capture all signal-responsive regulatory elements. In general, this strategy may be also useful to better understand the function of regulatory elements, especially as chromatin signatures help identify more candidate regulatory[50]. Since it is generally difficult to assign the function of regulatory elements when analyzing data from a steady-state sample, perturbing cells with various stimuli and identifying changes in H3K27ac can help begin to identify the function and gene targets of the expanding set of known regulatory elements in humans.

Finally, this study raises critical questions about the use of dynamic chromatin remodeling to identify signal-responsive regulatory elements: which chromatin marker(s) is/are most useful in identifying regulatory regions involved in signal-responsive gene expression change? Or are other chromatin proteins more reliable towards identifying signal-responsive enhancers? Previous studies have used dynamic changes, for example, in H3K4me2 and DNase I hypersensitivity[66, 133]. In theory, changes in p300 occupancy might be another strong candidate to analyze, because it binds tissue-specific enhancers[87]. However, at dynamic H3K27ac sites, the occupancy of H3K4me2 and H3 was stable, and the temporal change in p300 binding over the time course was uniform on average across H3K27ac variant sites. This may suggest H3K27ac is a better mark for identifying temporal changes in regulatory element activity. Yet, there was VEGFA-induced changes in DNase I hypersensitivity at enhancers identified using the HUVEC chromatin



states outside of H3K27ac variant sites. Comparatively, H3K27ac variant sites had no significant change in DNase I hypersensitivity over the time course. This suggests that a variety of chromatin marks and other factors change in response to an external stimulus but not always at the same signal-responsive regulatory elements. Hence, in order to capture all the signal-responsive regulatory elements for a particular signaling pathway, multiple different chromatin marks should be used.

## **5.2 p300 has Broad Effects on VEGFA-responsive Chromatin Remodeling**

p300 appears to have broad effects on H3K27ac. p300 bound a select number of sites within the candidate *cis* regulatory regions. Moreover, the binding of p300 appear to change temporally in one direction: an increase at Hour 1 before decreasing over the rest of the time course[150]. Yet, H3K27ac remodeling at H3K27ac variant sites were still sensitive to normal p300 activity even though the timing of H3K27ac change and p300 binding did not necessarily correlate. More specifically, at H0 and H4-12 sites, blocking p300 acetyltransferase activity inhibited H3K27ac changes at these sites even though p300 was strongest at these clusters in Hour 1. Additionally, VEGFA-responsive latent enhancers were not bound by p300 at any time point, but blocking p300 acetyltransferase activity altered H3K27ac occupancy at latent enhancers. Despite the temporal binding profile of p300, it appeared to indirectly regulate regulatory elements over time and in locations that lacked p300 binding. These broad effects of p300 may be accomplished through the regulating other chromatin modifiers, specifically other histone acetyltransferases (HATs) and histone deacetylases (HDACs), that can regulate H3K27ac remodeling outside of p300 binding. In particular, this could be accomplished by p300 acetylating such chromatin remodelers in response to VEGFA to regulate their activity. This would explain how H0, H4-12 and latent enhancers see p300-required changes in their H3K27ac profiles when p300 may not strongly bind these sites in position or in time. This could be tested by pulling down protein interacting with p300 via ChIP followed by mass spectrometry per time point. Mass spectrometry could identify candidate chromatin remodelers, likely HATs and HDACs, interacting with p300. This hypothesis would be further validated by analyzing temporal binding profiles of candidate chromatin remodelers via ChIP-seq, testing whether p300 acetylates these proteins in response to VEGFA and blocking their expression via siRNA to see if it blocks normal angiogenesis. An alternative explanation is that CBP has a complementary role to p300 in this time course, which would require further ChIP-seq experiments against CBP to resolve the binding of CBP in response to VEGFA

## **5.3 Stimulus-Responsive Genes Have Structures Relating to Dynamic Gene Expression Change**

Genes in the human genome have are diverse in size and structure[178], but how this structural diversity links with gene expression is poorly understood. One study previously re-

ported that primary responsive genes in immune cells in mouse tended to be smaller on average with fewer exons[217]. However, this study was performed prior to the development of gene expression profiling by RNA-seq and had a shorter time course[217] compared to this study. Consistent with the previous observation, VEGFA early responsive genes (ERGs) had a shorter gene length on average than any other gene set. Additionally, ERGs tended to have fewer exons per gene on average. Since ERGs were rapidly upregulated in response to VEGFA, the smaller size and fewer exons may aid rapid upregulation of expression by reducing the time Pol II spends in elongation, especially by reducing the number of pauses for co-transcriptional splicing. Comparatively, the other HUVEC gene sets, late responsive, downregulated response and non-responsive genes (LRGs, DRGs and NRGs, respectively) tended to have a similar gene size on average. Compared with ERGs, these gene sets may have less pressure to be smaller to enable rapid activation upon stimulation. Additionally, all three gene sets had a higher number of exons per gene on average, suggesting that these genes may be more liable to undergo alternative splicing. Surprisingly, NRGs tended to have more annotated exons on average than VEGFA-responsive genes. This suggests that NRGs have multiple isoforms, where multiple isoforms of NRGs could be expressed in this HUVEC time series. The current RNA-seq protocol of this study lacks the resolution to better understand any VEGFA-induced changes in isoform quantity, requiring further study to understand whether alternative splicing may mediate part of the VEGFA response.

## 5.4 The Regulation of H2A.Z Deposition by Promoter-Proximal Pol II Pausing

This study showed that H2A.Z deposition at a gene's TSS in mammals was strongly tied with the gene's PI. Surprisingly, the observed the correlation between these two factors was positive, which contrasted with previous studies in *Drosophila* where H2A.Z negatively correlate with paused Pol II build-up[41, 209]. This initially suggested that H2A.Z may promote promoter-proximal Pol II pausing and act instead as a barrier to productive elongation. However, the function of H2A.Z was found to be relatively consistent with previous studies[41, 42, 209]. When H2A.Z was knocked down via siRNA, it caused a global increase in promoter-proximal Pol II pausing, suggesting that the loss of H2A.Z incorporation into the nucleosome impedes the release of Pol II from pausing as in *Drosophila*[41]. It is unclear whether the observed increase in Pol II pausing was from a more stable nucleosome blocking Pol II elongation or that loss of H2A.Z inhibited recruitment of the MLL complex[42] to release it from its paused state. In the other direction, inhibiting Pol II pausing release increased deposition of H2A.Z at TSS, supporting that these two factors are tied together. Since H2A.Z is strongly conserved across species[36], this may suggest that complex that incorporates H2A.Z into the nucleosomes at TSS differs between species causing differences in the observed H2A.Z incorporation. Although this may explain the phenomenon, it remains unclear why the deposition of H2A.Z at TSS is different between human and *Drosophila* with respect to promoter-proximal Pol II pausing. Assuming changes in the complex that regulates H2A.Z deposition causes this effect, then the can-

didate gene for that complex from humans could be inserted into a transgenic *Drosophila* to test whether this alters H2A.Z deposition at TSS in *Drosophila* and, in turn, affects gene expression patterns.

## 5.5 Implications for Understanding the Regulation of Angiogenesis

This study gives a more comprehensive view of VEGFA-responsive gene expression relative to older microarray studies[159]. While this RNA-seq data strongly correlated with previous microarray data sets across multiple time points[150, 159], it also identified a variety of non-coding RNA genes temporally co-expressed with protein coding genes. Although non-coding RNAs have been implicated in endothelial cell function previously[218], this identifies almost 1,000 non-coding RNAs within endothelial cell function, about 300 of which are responsive to VEGFA. Further comparison with older microarray data sets may yield novel angiogenesis related protein-coding genes with the increased sensitivity of RNA-seq. Functionally, the differentially expressed protein-coding genes were consistent with the cellular phenotype of angiogenesis, although potential novel biological roles may lie within these VEGFA-responsive genes. Presumably, the co-differentially expressed non-coding RNA genes per VEGFA responsive gene cluster contribute towards the enriched functions of the protein coding genes, but further work with VEGFA-responsive non-coding RNAs is needed to understand their role in angiogenesis.

Pol II pausing release regulated several VEGFA responsive genes, but surprisingly increased pausing release appeared to only regulate a subset of VEGFA responsive genes. ERGs had an apparent, temporary increase in pausing release identified by a change in the PI distribution between Hours 0 and 1. None of the other gene sets had a similar significant change. This is supported by complementary report also studying transcriptional regulation in angiogenesis[181], although this study did not compare different sets of responsive genes. VEGFA may regulate rapidly responding genes through increased pausing release, but non-rapid gene expression changes may be regulated through changes in Pol II density. The necessity of pausing release was confirmed inhibiting HUVECs with FP and testing select ERGs, which FP inhibited rapid upregulation. Selected LRGs also were generally downregulated by FP treatment, but it is unclear whether the LRGs were affected due to loss of normal ERG expression or whether pausing release is also important for LRG expression change despite no significant change in PI on average. Curiously, it also appeared that tested ERGs started to increase in expression later in the time course under FP treatment, suggesting that pausing release may act as an amplifier but not the primary driver behind ERG expression change. Further studies are required to untangle the necessity of pausing release in mediating VEGFA-responsive gene expression change, especially by finding a way to block pausing release post-ERG expression to test the need for increased pausing release for LRG expression.

This study suggests several novel therapeutic targets for inhibiting angiogenesis. p300 inhibitors, such as C646, effectively blocked angiogenesis by inhibiting chromatin remodeling and responsive gene expression change[150]. Since p300 is important in many

cells, this may suggest that inhibiting other chromatin remodelers that mediate VEGFA-responsive regulatory regions, such as non-p300 HATs, may also be prime therapeutic targets. Blocking Pol II pausing release also negatively impacted angiogenesis-related gene expression, suggesting that FP or other related inhibitors, such as JQ1, may be useful. Several transcription factors assayed in this study appeared responsive to angiogenesis, any of which may be an effective therapeutic target, especially through siRNA. While all these strategies may inhibit angiogenesis, it is unclear how more effective these approaches would be against existing angiogenesis inhibitors, such as bevacizumab[155]. Inhibiting these potential therapeutic targets may negatively affect a wide range of other tissues within the body since many of these targets are present across many different cell types. Existing inhibitors often block the VEGFA molecule before reaching endothelial cells, which may allow greater specificity than these novel targets. However, recent advances in drug delivery nanotechnology have been able to deliver siRNAs specifically to endothelial cells in order to inhibit tumor angiogenesis[219]. With such advances, many of these pathways that regulate angiogenesis can be targeted without affecting other tissues if their delivery to endothelial cells is specific enough.

Finally, this study modeled the normal transcriptional regulation of sprouting angiogenesis, but it is unclear how this would be altered in endothelial cells undergoing angiogenesis caused by a disease process (*e.g.*, tumor). Notably, microarray studies suggest that tumor-associated endothelial cells undergoing angiogenesis have multiple differentially expressed genes compared with the gene expression profile in normal angiogenesis[220]. Disease-driven angiogenesis may cause a variety of gene expression changes in endothelial cells, which if identified could be utilized as novel therapeutic targets to specifically target the disease process. This analysis provides a baseline picture of the gene expression, chromatin and transcription changes that should occur, and further comparisons can be made against these profiles if repeated in endothelial cells derived from a disease process.



## Appendix A

# GWAS SNPs Mapping to HUVEC Enhancers and H3K27ac Variant Sites

Table A.1: Mapping GWAS SNPs to the H3K27ac variant sites (within a 2 kb window) shows that a number of GWAS SNPs map to these responsive enhancers, suggesting that such genotypic variation may affect the normal course of angiogenesis. Only SNPs with at least one hit at a H3K27ac variant site are shown here. Comparatively, the count of GWAS SNPs that hit all enhancer chromatin states is listed (without the 2 kb window), which includes any that mapped to the H3K27ac variant sites. Note, some rows may have more SNPs total within the H3K27ac variant clusters than within the enhancer states because some H3K27ac variant sites overlap promoters, which are not generally included in the enhancer chromatin state.

Disease Trait	H0	H1	H4-12	All Enhancer States
Height	8	6	2	64
Obesity-related traits	8	5	3	93
Bone mineral density	3	2	0	22
Crohn's disease	3	1	3	38
IgG glycosylation	2	16	6	59
Multiple sclerosis	2	3	0	24
HDL cholesterol	2	1	0	23
LDL cholesterol	2	1	0	16
Metabolite levels	2	1	0	5
Bipolar disorder and schizophrenia	2	0	2	12
Orofacial clefts	2	0	1	14
Androgen levels	2	0	0	2
Cholesterol, total	2	0	0	16
Femoral neck bone geometry and menarche (age at onset)	2	0	0	2
Migraine	2	0	0	11

Table A.1: (Continued)

<b>Disease Trait</b>	<b>H0</b>	<b>H1</b>	<b>H4-12</b>	<b>Total</b>
Response to antipsychotic treatment	2	0	0	5
Coronary heart disease	1	3	0	23
Body mass index	1	2	0	24
Ulcerative colitis	1	2	0	12
Schizophrenia	1	1	2	18
Bilirubin levels	1	1	1	10
Asthma	1	1	0	9
Immune response to anthrax vaccine	1	1	0	1
Primary biliary cirrhosis	1	1	0	6
Protein quantitative trait loci	1	1	0	12
Alzheimer's disease	1	0	2	4
Attention deficit hyperactivity disorder	1	0	2	12
Cognitive performance	1	0	2	10
Platelet counts	1	0	2	17
Serum dimethylarginine levels (symmetric)	1	0	2	6
Visceral adipose tissue/subcutaneous adipose tissue ratio	1	0	2	5
PR interval in <i>Tripanosoma cruzi</i> seropositivity	1	0	1	16
Adverse response to chemotherapy (neutropenia/leucopenia) (doxorubicin)	1	0	0	1
Alzheimer's disease (cognitive decline)	1	0	0	5
Blood pressure	1	0	0	3
Cerebrospinal T-tau levels	1	0	0	2
Chronic lymphocytic leukemia	1	0	0	8
Circulating vasoactive peptide levels	1	0	0	3
Diastolic blood pressure	1	0	0	0
DNA methylation (variation)	1	0	0	1
Graves' disease	1	0	0	2
Head circumference (infant)	1	0	0	1
Hypertension	1	0	0	0

Table A.1: (Continued)

<b>Disease Trait</b>	<b>H0</b>	<b>H1</b>	<b>H4-12</b>	<b>Total</b>
Insulin resistance/response	1	0	0	1
Liver enzyme levels (alanine transaminase)	1	0	0	3
Lung function (forced expiratory volume in 1 second)	1	0	0	2
Metabolic syndrome	1	0	0	9
Metabolite levels (5-HIAA/MHPG Ratio)	1	0	0	2
Metabolite levels (HVA/MHPG ratio)	1	0	0	2
Migraine without aura	1	0	0	3
MRI atrophy measures	1	0	0	2
Pancreatic cancer	1	0	0	2
Periodontitis (PAL4Q3)	1	0	0	2
Phospholipid levels (plasma)	1	0	0	9
Phosphorus levels	1	0	0	1
PR interval	1	0	0	4
Preeclampsia	1	0	0	1
Progressive supranuclear palsy	1	0	0	0
Proinsulin levels	1	0	0	0
Protein biomarker	1	0	0	1
Pubertal anthropometrics	1	0	0	2
Schizophrenia or bipolar disorder	1	0	0	1
Soluble levels of adhesion molecules	1	0	0	1
Stearic acid (18:0) plasma levels	1	0	0	1
Systolic blood pressure	1	0	0	1
Tonometry	1	0	0	3
Uric acid levels	1	0	0	3
Word reading	1	0	0	1
HIV-1 control	0	4	0	3
Triglycerides	0	4	0	22
Atopic dermatitis	0	2	0	8
Bipolar disorder	0	2	0	11
Chronic obstructive pulmonary disease-related biomarkers	0	2	0	4
Hair color	0	2	0	2

Table A.1: (Continued)

<b>Disease Trait</b>	<b>H0</b>	<b>H1</b>	<b>H4-12</b>	<b>Total</b>
Hematology traits	0	2	0	4
Melanoma	0	2	0	4
Multiple myeloma (IgH translocation)	0	2	0	4
Ovarian cancer	0	2	0	2
Response to antidepressant treatment	0	2	0	3
Response to serotonin reuptake inhibitors in major depressive disorder (plasma drug and metabolite levels)	0	2	0	1
Serum thyroid peroxidase antibody positivity	0	2	0	1
Waist Circumference - Triglycerides (WC-TG)	0	2	0	2
Breast cancer	0	1	3	21
Pulmonary function	0	1	3	10
Lentiform nucleus volume	0	1	2	3
White blood cell count	0	1	2	2
Age-related macular degeneration	0	1	1	8
Amyotrophic lateral sclerosis (sporadic)	0	1	1	9
Attention deficit hyperactivity disorder and conduct disorder	0	1	1	2
Interstitial lung disease	0	1	1	2
Lung function (forced expiratory flow between 25% and 75% of forced vital capacity)	0	1	1	1
Menarche (age at onset)	0	1	1	4
Pulmonary function (interaction)	0	1	1	10
Systemic sclerosis	0	1	1	3
Adverse response to chemotherapy (neutropenia/leucopenia) (all anthracycline-based drugs)	0	1	0	2
Adverse response to chemotherapy (neutropenia/leucopenia) (carboplatin)	0	1	0	3



Table A.1: (Continued)

<b>Disease Trait</b>	<b>H0</b>	<b>H1</b>	<b>H4-12</b>	<b>Total</b>
Adverse response to chemotherapy (neutropenia/leucopenia) (epirubicin)	0	1	0	1
AIDS progression	0	1	0	0
Ankylosing spondylitis	0	1	0	2
Asthma (childhood onset)	0	1	0	3
Axial length	0	1	0	2
Barrett's esophagus	0	1	0	3
Bipolar disorder (mood-incongruent)	0	1	0	2
Birth weight	0	1	0	4
Blond vs. brown hair color	0	1	0	1
Blood pressure measurement (high sodium intervention)	0	1	0	1
Bone mineral density (hip)	0	1	0	6
Bone mineral density (spine)	0	1	0	7
Chemerin levels	0	1	0	0
Complement C3 and C4 levels	0	1	0	0
Cutaneous nevi	0	1	0	3
D-dimer levels	0	1	0	2
Digestive system disease (Barrett's esophagus and esophageal adenocarcinoma combined)	0	1	0	3
Drug-induced liver injury (flucloxacillin)	0	1	0	0
Ejection fraction in <i>Tripanosoma cruzi</i> seropositivity	0	1	0	6
Esophageal adenocarcinoma	0	1	0	4
Eye color	0	1	0	2
Freckling	0	1	0	0
Glaucoma (exfoliation)	0	1	0	1
Hair morphology	0	1	0	3
Heart rate	0	1	0	5
Hematological and biochemical traits	0	1	0	5
Hippocampal atrophy	0	1	0	3
HIV-1 susceptibility	0	1	0	0

Table A.1: (Continued)

<b>Disease Trait</b>	<b>H0</b>	<b>H1</b>	<b>H4-12</b>	<b>Total</b>
Hypertension risk in short sleep duration	0	1	0	1
Hypothyroidism	0	1	0	3
IgG levels	0	1	0	0
Lung cancer	0	1	0	3
Lung Cancer (DNA repair capacity)	0	1	0	0
Major depressive disorder	0	1	0	8
Mean corpuscular hemoglobin	0	1	0	3
Mean corpuscular hemoglobin concentration	0	1	0	2
Mean corpuscular volume	0	1	0	3
Menopause (age at onset)	0	1	0	2
Narcolepsy with cataplexy	0	1	0	1
Other erythrocyte phenotypes	0	1	0	1
Parasitemia in <i>Trypanosoma cruzi</i> seropositivity	0	1	0	1
Periodontitis (Mean PAL)	0	1	0	2
Plasma amyloid beta peptide concentrations (ABx-42)	0	1	0	1
Prion diseases	0	1	0	2
Prostate cancer (gene x gene interaction)	0	1	0	5
Psoriasis	0	1	0	2
QRS duration in <i>Trypanosoma cruzi</i> seropositivity	0	1	0	2
Response to amphetamines	0	1	0	8
Response to angiotensin II receptor blocker therapy	0	1	0	2
Response to statin therapy	0	1	0	3
Response to Vitamin E supplementation	0	1	0	1
Serum protein levels (sST2)	0	1	0	7
Sickle cell anemia (haemolysis)	0	1	0	1
Smoking quantity	0	1	0	1
Smooth-surface caries	0	1	0	7
Testosterone levels	0	1	0	1

Table A.1: (Continued)

<b>Disease Trait</b>	<b>H0</b>	<b>H1</b>	<b>H4-12</b>	<b>Total</b>
Thoracic aortic aneurysms and dissections	0	1	0	0
Thyroid function	0	1	0	0
Tumor biomarkers	0	1	0	0
Type 1 diabetes	0	1	0	11
Urate levels	0	1	0	16
Vertical cup-disc ratio	0	1	0	1
Vitiligo	0	1	0	0
Waist circumference	0	1	0	4
Prostate cancer	0	0	5	11
Adiponectin levels	0	0	2	12
Alcohol consumption	0	0	2	6
Body mass index (interaction)	0	0	2	2
Breast cancer (early onset)	0	0	2	3
Folate pathway vitamin levels	0	0	2	3
Quantitative traits	0	0	2	7
Serum dimethylarginine levels (asymmetric/symmetric ratio)	0	0	2	9
Type 2 diabetes	0	0	2	28
Airflow obstruction	0	0	1	6
Amyotrophic lateral sclerosis (age of onset)	0	0	1	4
Biomedical quantitative traits	0	0	1	2
Brain structure	0	0	1	6
Breast size	0	0	1	5
C-reactive protein	0	0	1	15
Caffeine consumption	0	0	1	2
Circulating myeloperoxidase levels (serum)	0	0	1	1
Coffee consumption	0	0	1	3
Conduct disorder	0	0	1	1
Coronary artery calcification	0	0	1	6
Creatinine levels	0	0	1	0
Gallbladder cancer	0	0	1	2
Homeostasis model assessment of beta-cell function (interaction)	0	0	1	2
Illicit drug use	0	0	1	0
Inflammatory biomarkers	0	0	1	5

Table A.1: (Continued)

<b>Disease Trait</b>	<b>H0</b>	<b>H1</b>	<b>H4-12</b>	<b>Total</b>
Inflammatory bowel disease	0	0	1	15
Left ventricular mass	0	0	1	1
Liver enzyme levels (gamma-glutamyl transferase)	0	0	1	5
Major depressive disorder (broad)	0	0	1	2
Mean forced vital capacity from 2 exams	0	0	1	1
Metabolite levels (Pyroglutamine)	0	0	1	5
Multiple sclerosis–Brain Glutamate Levels	0	0	1	2
PCA3 expression level	0	0	1	1
Periodontitis (DPAL)	0	0	1	1
Personality dimensions	0	0	1	3
Post-traumatic stress disorder	0	0	1	1
Prostate-specific antigen levels	0	0	1	3
Psychosis (atypical)	0	0	1	3
Pulmonary function decline	0	0	1	5
Reading and spelling	0	0	1	2
Red blood cell traits	0	0	1	5
Response to mTOR inhibitor (everolimus)	0	0	1	2
Response to protease inhibitor treatment in hepatitis c (peak serum total bilirubin levels)	0	0	1	1
Retinal arteriolar caliber	0	0	1	1
Retinopathy in non-diabetics	0	0	1	2
Rheumatoid arthritis	0	0	1	25
Serum prostate-specific antigen levels	0	0	1	1
Serum total protein level	0	0	1	2
Smoking behavior	0	0	1	9
Subcutaneous adipose tissue	0	0	1	2
Sudden cardiac arrest	0	0	1	4
Systemic lupus erythematosus	0	0	1	13
Systolic blood pressure (alcohol consumption interaction)	0	0	1	3



Table A.1: (Continued)

<b>Disease Trait</b>	<b>H0</b>	<b>H1</b>	<b>H4-12</b>	<b>Total</b>
Total ventricular volume	0	0	1	4
Venous thromboembolism (gene x gene interaction)	0	0	1	3
Visceral adipose tissue ad- justed for BMI	0	0	1	4
White blood cell types	0	0	1	2

# Appendix B

## Additional Tables and Figures For Promoter-Proximal Pol II Pausing Analysis

### B.1 Tables

Table B.1: Summary of human and mouse cell lines used in Chapter 4.

<b>Name/Abbreviation</b>	<b>Description</b>	<b>Human or Mouse</b>
H1	Human Embryonic Stem Cell	Human
IMR90	Fetal Lung Fibroblast	Human
MCF7	Breast Cancer Cell Line	Human
GM12878	Epstein-Barr Immortalized B-cell	Human
GM12891	Epstein-Barr Immortalized B-cell	Human
GM12892	Epstein-Barr Immortalized B-cell	Human
MM1.S	Multiple myeloma	Human
U87	Glioblastoma	Human
K562	Leukemia	Human
HUVEC	Human Umbilical Vascular Endothelial Cell	Human
HepG2	Heptocellular Carcinoma	Human
H128	Small Cell Lung Cancer	Human
H2171	Small Cell Lung Cancer	Human
A549	Epithelial Cell Line, Lung Carcinoma Tissue	Human
P493	B cell model of Burkitt's lymphoma	Human
HeLa	Cervical Carcinoma	Human
CH12	B-cell lymphoma	Mouse
Bone Marrow	Bone Marrow, Adult	Mouse
BMDMacrophage	Bone Marrow Derived Macrophage	Mouse
Cerebellum	Cerebellum, Adult	Mouse
Cortex	Cortex, Adult	Mouse

Table B.1: Summary of human and mouse cell lines used in Chapter 4.

<b>Name/Abbreviation</b>	<b>Description</b>	<b>Human or Mouse</b>
Heart	Heart, Adult	Mouse
Kidney	Kidney, Adult	Mouse
Limb	Limb, Fetal	Mouse
Liver	Liver, Adult	Mouse
Lung	Lung, Adult	Mouse
MEF	Embryonic Fibroblast	Mouse
MEL	Leukemia	Mouse
Olfactory	Olfactory, Adult	Mouse
Small Intestine	Adult 8wks, Small Intestine	Mouse
Spleen	Spleen, Adult	Mouse
Testis	Testis, Adult	Mouse
Thymus	Thymus, Adult	Mouse
Whole Brain	Whole Brain, Fetal	Mouse
mES	Embryonic Stem Cell	Mouse

## **B.2 Figures**

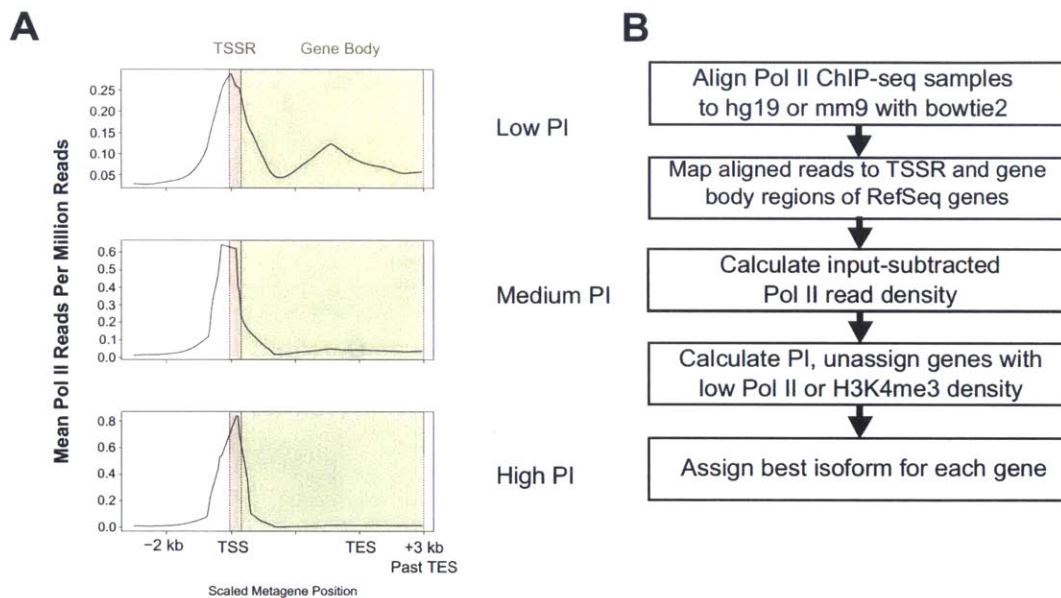


Figure B-1: In A, smoothed metagene profiles showing Pol II averages of genes divided into the top, middle, and bottom quantiles by PI in GM12878. TSSR, the transcriptional start site (TSS) region, was defined as TSS -50 to +300 bp. Gene body was defined as TSS+300 bp to transcriptional end site (TES) +3 kb. In B, the workflow for processing Pol II ChIP-seq data and assigning a PI value to each gene in the genome with sufficient RNAP2 or H3K4me3 density at the TSS.

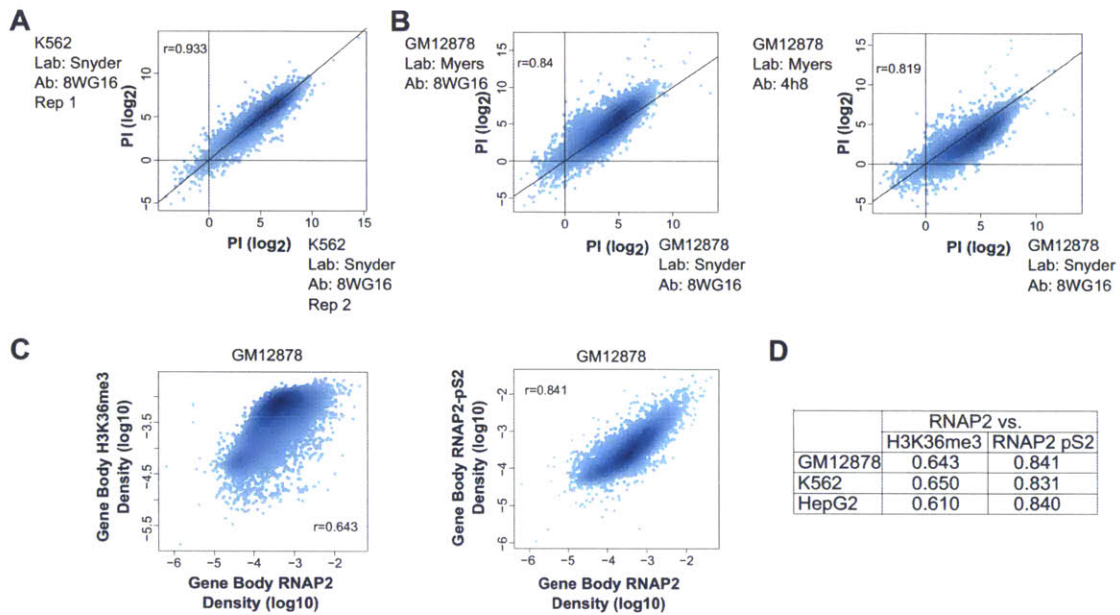


Figure B-2: In A-B, comparison of Pol II PI calling across biological replicates from the same and different labs and across different Pol II antibodies. In A, biological replicates of the K562 cell line from the same lab with the same antibody. In B, biological replicates of the GM12878 cell line from different labs using the same antibody, or from different labs using different antibodies. In C-D, gene body Pol II density correlation with H3K36me3 and Pol II pS2 across samples. Pearson correlations are shown throughout this figure.



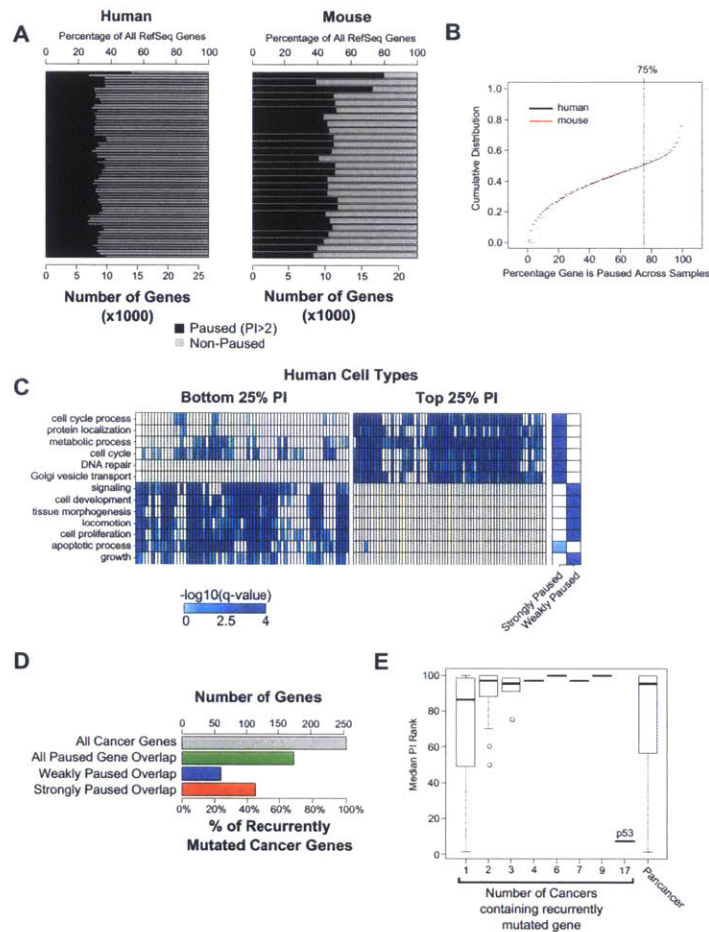


Figure B-3: In A, extended distribution of paused and non-paused genes across all analyzed Pol II ChIP-seq samples (expands Fig. 1b, first two rows of each set are “Total” and “Shared”). In B, empirical cumulative distribution of the percentage of samples analyzed in which a gene is paused. The majority of paused genes are paused in 75% of samples or more. In C, extended GO term analysis of the genes with the top and bottom quartile of PI in individual human samples, and of genes strongly or weakly paused across all human samples. In D, number and percentage of Lawrence *et al.* (2014) recurrently mutated cancer genes that overlap human paused genes. Notably the majority of recurrently mutated genes in this list were found among paused genes in our human samples. In E, genes recurrently mutated in increasingly more cancer types in Lawrence *et al.* (2014) had increasing median PI rank until frequency 17 (all samples). The only gene recurrently mutated in more than 9 samples was p53, which had a low median PI rank (bottom 15%). Genes recurrently mutated when considering the union of all samples (“pancancer”) were more highly paused on average.

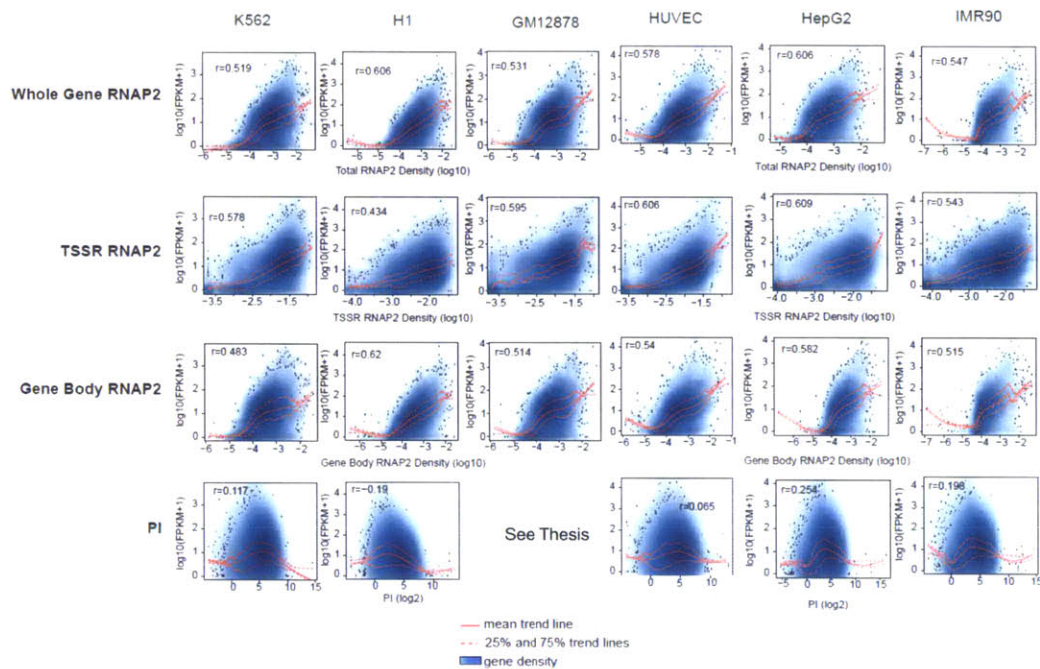


Figure B-4: The trend between each indicated parameter and gene expression level was calculated. Overall, we consistently observed that Pol II density best correlation with gene expression levels whereas the PI correlated less well. Furthermore, we observed a consistent “hill-shape” relationship between gene expression and PI, suggesting that within a cell type genes with relatively high and low PI values tended to be not as strongly expressed as genes with intermediate PIs.

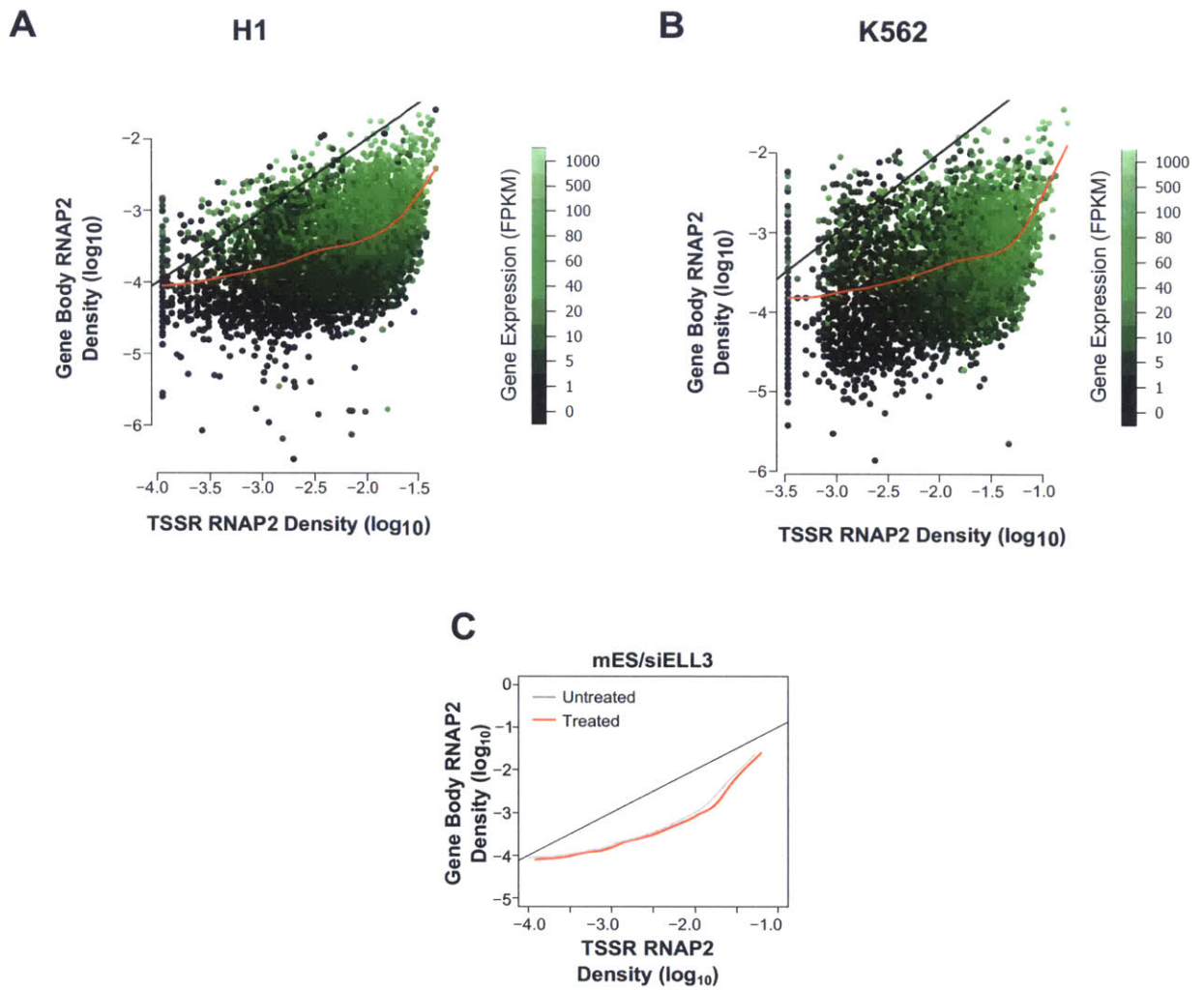


Figure B-5: A-B. Visualization of data from H1 and K562 cells. C. siRNA knockdown of ELL3 in murine ES cells had minimal effect on the TSSR-gene body Pol II density trend line (black line is where PI = 1).

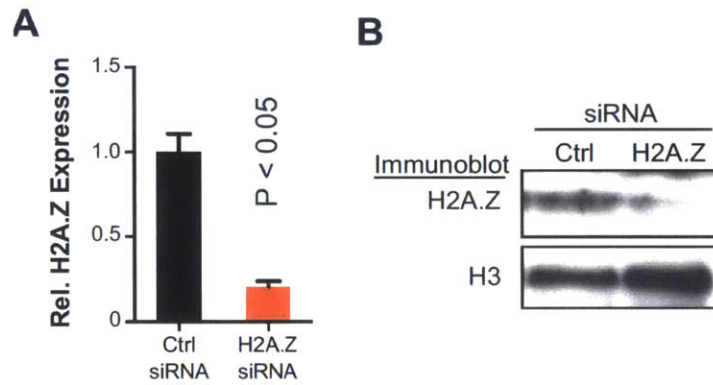


Figure B-6: In A, quantitation of H2A.Z mRNA levels by qRT-PCR showed significant knockdown by H2A.Z siRNA. In B, immunoblot showing H2A.Z protein depletion by siRNA transfection of MCF7 cells. Histone H3 was used as a loading control. (Data and result provided by Dr. Bing Zhang.)

# References

1. Carmeliet, P. Angiogenesis in health and disease. *Nature Medicine* **9**, 653–660 (June 2003).
2. Carmeliet, P. Mechanisms of angiogenesis and arteriogenesis. *Nature Medicine* **6**, 389–395 (Apr. 2000).
3. Phng, L.-K. & Gerhardt, H. Angiogenesis: A Team Effort Coordinated by Notch. *Developmental Cell* **16**, 196–208. ISSN: 1534-5807 (Feb. 2009).
4. Eilken, H. M. & Adams, R. H. Dynamics of endothelial cell behavior in sprouting angiogenesis. *Current Opinion in Cell Biology* **22**, 617–625. ISSN: 0955-0674 (Oct. 2010).
5. De Val, S. & Black, B. L. Transcriptional Control of Endothelial Cell Development. *Developmental Cell* **16**, 180–195. ISSN: 1534-5807 (Feb. 2009).
6. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921. ISSN: 0028-0836 (Feb. 2001).
7. Gerstein, M. B. *et al.* What is a gene, post-ENCODE? History and updated definition. en. *Genome Research* **17**, 669–681. ISSN: 1088-9051, 1549-5469 (June 2007).
8. Zhou, V. W., Goren, A. & Bernstein, B. E. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* **12**, 7–18. ISSN: 1471-0056 (Jan. 2011).
9. Furey, T. S. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews Genetics* **13**, 840–852. ISSN: 1471-0056 (Dec. 2012).
10. Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**, 669–680. ISSN: 1471-0056 (Oct. 2009).
11. Crick, F. Central Dogma of Molecular Biology. en. *Nature* **227**, 561–563 (Aug. 1970).
12. Kauffman, S. A. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology* **22**, 437–467. ISSN: 0022-5193 (Mar. 1969).
13. Consortium, I. H. G. S. Finishing the euchromatic sequence of the human genome. en. *Nature* **431**, 931–945. ISSN: 0028-0836 (Oct. 2004).
14. Ong, C.-T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet. Modes of Transcription* **12**, 283–293. ISSN: 1471-0056 (Apr. 2011).



15. Ong, C.-T. & Corces, V. G. Enhancers: emerging roles in cell fate specification. *EMBO reports* **13**, 423–430 (May 2012).
16. Consortium, T. E. P. An integrated encyclopedia of DNA elements in the human genome. en. *Nature* **489**, 57–74. ISSN: 0028-0836 (Sept. 2012).
17. Graur, D. *et al.* On the Immortality of Television Sets: “Function” in the Human Genome According to the Evolution-Free Gospel of ENCODE. en. *Genome Biology and Evolution* **5**, 578–590. ISSN: , 1759-6653 (Jan. 2013).
18. Rands, C. M., Meader, S., Ponting, C. P. & Lunter, G. 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage. *PLoS Genet* **10**, e1004525 (July 2014).
19. Alberts, B. *et al.* en. in *Molecular Biology of the Cell* 4th Edition, Chapter 6 (Garland Science, New York, 2002). (Visited on 05/20/2015).
20. Kornblihtt, A. R. *et al.* Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature Reviews Molecular Cell Biology* **14**, 153–165. ISSN: 1471-0072 (Mar. 2013).
21. Bentley, D. L. Coupling mRNA processing with transcription in time and space. en. *Nature Reviews Genetics* **15**, 163–175. ISSN: 1471-0056 (Feb. 2014).
22. Ameur, A. *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. en. *Nature Structural & Molecular Biology* **18**, 1435–1440. ISSN: 1545-9993 (Dec. 2011).
23. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476. ISSN: 0028-0836 (Nov. 2008).
24. Djebali, S. *et al.* Landscape of transcription in human cells. en. *Nature* **489**, 101–108. ISSN: 0028-0836 (Sept. 2012).
25. Rinn, J. L. & Chang, H. Y. Genome Regulation by Long Noncoding RNAs. *Annual Review of Biochemistry* **81**, 145–166 (2012).
26. Eddy, S. R. Non-coding RNA genes and the modern RNA world. *Nature Reviews Genetics* **2**, 919–929. ISSN: 1471-0056 (Dec. 2001).
27. He, L. & Hannon, G. J. MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews Genetics* **5**, 522–531. ISSN: 1471-0056 (July 2004).
28. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227. ISSN: 0028-0836 (Mar. 2009).
29. Guttman, M. *et al.* lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295–300. ISSN: 1476-4687 (Sept. 2011).
30. Alberts, B. *et al.* en. in *Molecular Biology of the Cell* 4th Edition, Chapter 4 (Garland Science, New York, 2002). (Visited on 05/15/2015).
31. Cavalli, G. & Misteli, T. Functional implications of genome topology. en. *Nature Structural & Molecular Biology* **20**, 290–299. ISSN: 1545-9993 (Mar. 2013).

32. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**, 289–293 (Oct. 2009).
33. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Research* **21**, 381–395. ISSN: 1001-0602 (Mar. 2011).
34. Elgin, S. C. Heterochromatin and gene regulation in *Drosophila*. *Current Opinion in Genetics & Development* **6**, 193–202. ISSN: 0959-437X (Apr. 1996).
35. Babu, A. & Verma, R. S. en. in *International Review of Cytology* 1–49 (Academic Press, Oct. 1987). ISBN: 9780080586458.
36. Talbert, P. B. & Henikoff, S. Histone variants – ancient wrap artists of the epigenome. *Nature Reviews Molecular Cell Biology* **11**, 264–275. ISSN: 1471-0072 (Apr. 2010).
37. Harshman, S. W., Young, N. L., Parthun, M. R. & Freitas, M. A. H1 histones: current perspectives and challenges. en. *Nucleic Acids Research* **41**, 9593–9609. ISSN: 0305-1048, 1362-4962 (Nov. 2013).
38. Berman, H. M. *et al.* The Protein Data Bank. en. *Nucleic Acids Research* **28**, 235–242. ISSN: 0305-1048, 1362-4962 (Jan. 2000).
39. Darekk2. *Nucleosome Crystal Structure* <[http://en.wikipedia.org/wiki/File:Nucleosome\\_core\\_particle\\_1EQZ\\_large.gif](http://en.wikipedia.org/wiki/File:Nucleosome_core_particle_1EQZ_large.gif)>.
40. Kossel, A. *Ueber die chemische Beschaffenheit des Zellkerns...* de (P. A. Norstedt, 1911).
41. Weber, C., Ramachandran, S. & Henikoff, S. Nucleosomes Are Context-Specific, H2A.Z-Modulated Barriers to RNA Polymerase. *Molecular Cell* **53**, 819–830. ISSN: 1097-2765 (Mar. 2014).
42. Hu, G. *et al.* H2A.Z Facilitates Access of Active and Repressive Complexes to Chromatin in Embryonic Stem Cell Self-Renewal and Differentiation. *Cell Stem Cell* **12**, 180–192. ISSN: 1934-5909 (Feb. 2013).
43. Tolstorukov, M. Y., Kharchenko, P. V., Goldman, J. A., Kingston, R. E. & Park, P. J. Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes. en. *Genome Research* **19**, 967–977. ISSN: 1088-9051, 1549-5469 (June 2009).
44. Kouzarides, T. Chromatin Modifications and Their Function. *Cell* **128**, 693–705. ISSN: 0092-8674 (Feb. 2007).
45. Vidali, G., Gershey, E. L. & Allfrey, V. G. Chemical Studies of Histone Acetylation THE DISTRIBUTION OF  $\hat{\text{I}}\text{-N-ACETYLLYSINE}$  IN CALF THYMUS HISTONES. en. *Journal of Biological Chemistry* **243**, 6361–6366. ISSN: 0021-9258, 1083-351X (Dec. 1968).
46. Kayne, P. S. *et al.* Extremely conserved histone H4 N terminus is dispensable for growth but essential for repressing the silent mating loci in yeast. *Cell* **55**, 27–39. ISSN: 0092-8674 (Oct. 1988).

47. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45. ISSN: 0028-0836 (Jan. 2000).
48. Jenuwein, T. & Allis, C. D. Translating the Histone Code. en. *Science* **293**, 1074–1080. ISSN: 0036-8075, 1095-9203 (Aug. 2001).
49. Kharchenko, P. V. *et al.* Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* **471**, 480–485. ISSN: 0028-0836 (Mar. 2011).
50. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49. ISSN: 0028-0836 (May 2011).
51. Vo, N. & Goodman, R. H. CREB-binding Protein and p300 in Transcriptional Regulation. en. *Journal of Biological Chemistry* **276**, 13505–13508. ISSN: 0021-9258, 1083-351X (Apr. 2001).
52. Collas, P. The Current State of Chromatin Immunoprecipitation. en. *Molecular Biotechnology* **45**, 87–100. ISSN: 1073-6085, 1559-0305 (Jan. 2010).
53. Rhee, H. S. & Pugh, B. F. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell* **147**, 1408–1419. ISSN: 0092-8674 (Dec. 2011).
54. Steensel, B. v. & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered Dam methyltransferase. *Nature Biotechnology* **18**, 424–428. ISSN: 1087-0156 (Apr. 2000).
55. Buck, M. J. & Lieb, J. D. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83**, 349–360. ISSN: 0888-7543 (Mar. 2004).
56. Jackson, V. Studies on histone organization in the nucleosome using formaldehyde as a reversible cross-linking agent. *Cell* **15**, 945–954. ISSN: 0092-8674 (Nov. 1978).
57. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. en. *Science* **270**, 467–470. ISSN: 0036-8075, 1095-9203 (Oct. 1995).
58. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112. ISSN: 0028-0836 (May 2009).
59. Jon. *ChIP-seq* <[http://en.wikipedia.org/wiki/ChIP-sequencing#mediaviewer/File:Chromatin\\_immunoprecipitation\\_sequencing.svg](http://en.wikipedia.org/wiki/ChIP-sequencing#mediaviewer/File:Chromatin_immunoprecipitation_sequencing.svg)>.
60. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. en. *Nature Methods* **9**, 357–359. ISSN: 1548-7091 (Apr. 2012).
61. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. en. *Bioinformatics* **25**, 1754–1760. ISSN: 1367-4803, 1460-2059 (July 2009).
62. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, R137 (2008).

63. Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotech* **26**, 1351–1359. ISSN: 1087-0156 (Dec. 2008).
64. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**, 823–837. ISSN: 0092-8674 (May 2007).
65. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics* **39**, 311–318. ISSN: 1061-4036 (Mar. 2007).
66. He, H. H. *et al.* Nucleosome dynamics define transcriptional enhancers. en. *Nature Genetics* **42**, 343–347. ISSN: 1061-4036 (Mar. 2010).
67. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences* **107**, 21931–21936 (Dec. 2010).
68. Roh, T.-Y., Cuddapah, S., Cui, K. & Zhao, K. The genomic landscape of histone modifications in human T cells. en. *Proceedings of the National Academy of Sciences* **103**, 15782–15787. ISSN: 0027-8424, 1091-6490 (Oct. 2006).
69. Wagner, E. J. & Carpenter, P. B. Understanding the language of Lys36 methylation at histone H3. *Nature Reviews Molecular Cell Biology* **13**, 115–126. ISSN: 1471-0072 (Feb. 2012).
70. Adam, M., Robert, F., Larochelle, M. & Gaudreau, L. H2A.Z Is Required for Global Chromatin Integrity and for Recruitment of RNA Polymerase II under Specific Conditions. en. *Molecular and Cellular Biology* **21**, 6270–6279. ISSN: 0270-7306, 1098-5549 (Sept. 2001).
71. Schones, D. E. *et al.* Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell* **132**, 887–898. ISSN: 0092-8674 (Mar. 2008).
72. Boyle, A. P. *et al.* High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* **132**, 311–322. ISSN: 0092-8674 (Jan. 2008).
73. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. en. *Nature* **489**, 75–82. ISSN: 0028-0836 (Sept. 2012).
74. Ruthenburg, A. J., Allis, C. D. & Wysocka, J. Methylation of Lysine 4 on Histone H3: Intricacy of Writing and Reading a Single Epigenetic Mark. English. *Molecular Cell* **25**, 15–30. ISSN: 1097-2765 (Dec. 2007).
75. KarliÅĀ, R., Chung, H.-R., Lasserre, J., VlahoviÅĀek, K. & Vingron, M. Histone modification levels are predictive for gene expression. en. *Proceedings of the National Academy of Sciences* **107**, 2926–2931. ISSN: 0027-8424, 1091-6490 (Feb. 2010).
76. Day, D. S., Luquette, L. J., Park, P. J. & Kharchenko, P. V. Estimating enrichment of repetitive elements from high-throughput sequence data. *Genome Biology* **11**, R69. ISSN: 1465-6906 (2010).

77. Adelman, K. & Lis, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature Reviews Genetics* **13**, 720–731. ISSN: 1471-0056 (Oct. 2012).
78. Lauberth, S. *et al.* H3K4me3 Interactions with TAF3 Regulate Preinitiation Complex Assembly and Selective Gene Activation. *Cell* **152**, 1021–1036. ISSN: 0092-8674 (Feb. 2013).
79. Benayoun, B. *et al.* H3K4me3 Breadth Is Linked to Cell Identity and Transcriptional Consistency. *Cell* **158**, 673–688. ISSN: 0092-8674 (July 2014).
80. Pekowska, A. *et al.* H3K4 tri-methylation provides an epigenetic signature of active enhancers. *The EMBO Journal* **30**, 4198–4210 (Oct. 2011).
81. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187. ISSN: 0028-0836 (May 2010).
82. Bonn, S. *et al.* Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat Genet* **44**, 148–156. ISSN: 1546-1718 (Jan. 2012).
83. Kowalczyk, M. S. *et al.* Intragenic Enhancers Act as Alternative Promoters. English. *Molecular Cell* **45**, 447–458. ISSN: 1097-2765 (Feb. 2012).
84. Pekowska, A., Benoukraf, T., Ferrier, P. & Spicuglia, S. A unique H3K4me2 profile marks tissue-specific gene regulation. *en. Genome Research* **20**, 1493–1502. ISSN: 1088-9051, 1549-5469 (Nov. 2010).
85. Ostuni, R. *et al.* Latent Enhancers Activated by Stimulation in Differentiated Cells. *Cell* **152**, 157–171. ISSN: 0092-8674 (Jan. 2013).
86. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics* **13**, 613–626. ISSN: 1471-0056 (Sept. 2012).
87. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858. ISSN: 0028-0836 (Feb. 2009).
88. Kamakaka, R. T. & Biggins, S. Histone variants: deviants? *en. Genes & Development* **19**, 295–316. ISSN: 0890-9369, 1549-5477 (Feb. 2005).
89. Bargaje, R. *et al.* Proximity of H2A.Z containing nucleosome to the transcription start site influences gene expression levels in the mammalian liver and brain. *en. Nucleic Acids Research* **40**, 8965–8978. ISSN: 0305-1048, 1362-4962 (Oct. 2012).
90. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560. ISSN: 0028-0836 (Aug. 2007).
91. Gifford, C. A. *et al.* Transcriptional and Epigenetic Dynamics during Specification of Human Embryonic Stem Cells. *Cell* **153**, 1149–1163. ISSN: 0092-8674 (May 2013).
92. Xie, W. *et al.* Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells. *Cell* **153**, 1134–1148. ISSN: 0092-8674 (May 2013).



93. Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. en. *Nature* **469**, 343–349. ISSN: 0028-0836 (Jan. 2011).
94. Bernstein, B. E. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125**, 315–326. ISSN: 0092-8674 (Apr. 2006).
95. Djupedal, I. & Ekwall, K. Epigenetics: heterochromatin meets RNAi. *Cell Research* **19**, 282–295. ISSN: 1001-0602 (2009).
96. Ho, J. W. K. *et al.* Comparative analysis of metazoan chromatin organization. en. *Nature* **512**, 449–452. ISSN: 0028-0836 (Aug. 2014).
97. Kim, S., Kim, H., Fong, N., Erickson, B. & Bentley, D. L. Pre-mRNA splicing is a determinant of histone H3K36 methylation. en. *Proceedings of the National Academy of Sciences* **108**, 13564–13569. ISSN: 0027-8424, 1091-6490 (Aug. 2011).
98. Wen, H. *et al.* ZMYND11 links histone H3.3K36me3 to transcription elongation and tumour suppression. en. *Nature* **508**, 263–268. ISSN: 0028-0836 (Apr. 2014).
99. Gross, D. S. & Garrard, W. T. Nuclease Hypersensitive Sites in Chromatin. *Annual Review of Biochemistry* **57**, 159–197 (1988).
100. Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M. & Teichmann, S. A. Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology* **14**, 283–291. ISSN: 0959-440X (June 2004).
101. Brivanlou, A. H. & Darnell, J. E. Signal Transduction and the Control of Gene Expression. en. *Science* **295**, 813–818. ISSN: 0036-8075, 1095-9203 (Feb. 2002).
102. Sawadogo, M & Sentenac, A. RNA Polymerase B (II) and General Transcription Factors. *Annual Review of Biochemistry* **59**, 711–754 (1990).
103. Sandelin, A., Alkema, W., Engstr m, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open access database for eukaryotic transcription factor binding profiles. en. *Nucleic Acids Research* **32**, D91–D94. ISSN: 0305-1048, 1362-4962 (Jan. 2004).
104. Bailey, T. L., Williams, N., Misleh, C. & Li, W. W. MEME: discovering and analyzing DNA and protein sequence motifs. en. *Nucleic Acids Research* **34**, W369–W373. ISSN: 0305-1048, 1362-4962 (July 2006).
105. Bailey, T. L. DREME: motif discovery in transcription factor ChIP-seq data. en. *Bioinformatics* **27**, 1653–1659. ISSN: 1367-4803, 1460-2059 (June 2011).
106. Machanick, P. & Bailey, T. L. MEME-ChIP: Motif Analysis of Large DNA Datasets. en. *Bioinformatics* **27**, 1696–1697. ISSN: 1367-4803, 1460-2059 (June 2011).
107. Oliphant, A. R., Brandl, C. J. & Struhl, K. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. en. *Molecular and Cellular Biology* **9**, 2944–2949. ISSN: 0270-7306, 1098-5549 (July 1989).
108. Berger, M. F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology* **24**, 1429–1435. ISSN: 1087-0156 (Nov. 2006).

109. Hollenhorst, P. C., McIntosh, L. P. & Graves, B. J. Genomic and Biochemical Insights into the Specificity of ETS Transcription Factors. *Annual Review of Biochemistry* **80**, 437–471 (2011).
110. Sharrocks, A. D. The ETS-domain transcription factor family. *Nature Reviews Molecular Cell Biology* **2**, 827–837. ISSN: 1471-0072 (Nov. 2001).
111. Lin, C. *et al.* Transcriptional Amplification in Tumor Cells with Elevated c-Myc. *Cell* **151**, 56–67. ISSN: 0092-8674 (Sept. 2012).
112. Wolf, E., Lin, C. Y., Eilers, M. & Levens, D. L. Taming of the beast: shaping Myc-dependent amplification. *Trends in Cell Biology* **25**, 241–248. ISSN: 0962-8924 (Apr. 2015).
113. Garber, M. *et al.* A High-Throughput Chromatin Immunoprecipitation Approach Reveals Principles of Dynamic Gene Regulation in Mammals. *Molecular Cell* **47**, 810–822. ISSN: 1097-2765 (Sept. 2012).
114. Xie, D. *et al.* Dynamic trans-Acting Factor Colocalization in Human Cells. *Cell* **155**, 713–724. ISSN: 0092-8674 (Oct. 2013).
115. Maston, G. A., Landt, S. G., Snyder, M. & Green, M. R. Characterization of Enhancer Function from Genome-Wide Analyses. *Annual Review of Genomics and Human Genetics* **13**, 29–57 (2012).
116. Whyte, W. *et al.* Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* **153**, 307–319. ISSN: 0092-8674 (Apr. 2013).
117. LovÅp'n, J. *et al.* Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers. *Cell* **153**, 320–334. ISSN: 0092-8674 (Apr. 2013).
118. Noonan, J. P. & McCallion, A. S. Genomics of Long-Range Regulatory Elements. *Annual Review of Genomics and Human Genetics* **11**, 1–23 (2010).
119. Merkenschlager, M. & Odom, D. CTCF and Cohesin: Linking Gene Regulatory Elements with Their Targets. *Cell* **152**, 1285–1297. ISSN: 0092-8674 (Mar. 2013).
120. Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. en. *Nature* **467**, 430–435. ISSN: 0028-0836 (Sept. 2010).
121. Allen, B. L. & Taatjes, D. J. The Mediator complex: a central integrator of transcription. en. *Nature Reviews Molecular Cell Biology* **16**, 155–166. ISSN: 1471-0072 (Mar. 2015).
122. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. en. *Nature* **485**, 376–380. ISSN: 0028-0836 (May 2012).
123. Sexton, T. *et al.* Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell* **148**, 458–472. ISSN: 0092-8674 (Feb. 2012).
124. Rao, S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680. ISSN: 0092-8674 (Dec. 2014).

125. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. en. *Nature* **470**, 279–283. ISSN: 0028-0836 (Feb. 2011).
126. M̄ajller, C. & Leutz, A. Chromatin remodeling in development and differentiation. *Current Opinion in Genetics & Development* **11**, 167–174. ISSN: 0959-437X (Apr. 2001).
127. Ho, L. & Crabtree, G. R. Chromatin remodelling during development. *Nature* **463**, 474–484. ISSN: 0028-0836 (Jan. 2010).
128. Zhu, J. *et al.* Genome-wide Chromatin State Transitions Associated with Developmental and Environmental Cues. *Cell* **152**, 642–654. ISSN: 0092-8674 (Jan. 2013).
129. Towbin, B. D., Gonzalez-Sandoval, A. & Gasser, S. M. Mechanisms of heterochromatin subnuclear localization. *Trends in Biochemical Sciences* **38**, 356–363. ISSN: 0968-0004 (July 2013).
130. Weiner, A. *et al.* Systematic Dissection of Roles for Chromatin Regulators in a Yeast Stress Response. *PLoS Biol* **10**, e1001369 (July 2012).
131. Malik, A. N. *et al.* Genome-wide identification and characterization of functional neuronal activity-dependent enhancers. en. *Nature Neuroscience* **17**, 1330–1339. ISSN: 1097-6256 (2014).
132. Fowler, T., Sen, R. & Roy, A. Regulation of Primary Response Genes. *Molecular Cell* **44**, 348–360. ISSN: 1097-2765 (Nov. 2011).
133. He, H. H. *et al.* Differential DNase I Hypersensitivity Reveals Factor-Dependent Chromatin Dynamics. en. *Genome Research* **22**, 1015–1025. ISSN: 1088-9051, 1549-5469 (2012).
134. Koike, N. *et al.* Transcriptional Architecture and Chromatin Landscape of the Core Circadian Clock in Mammals. en. *Science* **338**, 349–354. ISSN: 0036-8075, 1095-9203 (Oct. 2012).
135. Jonkers, I. & Lis, J. T. Getting up to speed with transcription elongation by RNA polymerase II. en. *Nature Reviews Molecular Cell Biology* **16**, 167–177. ISSN: 1471-0072 (2015).
136. Zhou, Q., Li, T. & Price, D. H. RNA Polymerase II Elongation Control. *Annual Review of Biochemistry* **81**, 119–143 (2012).
137. Herr, A. J., Jensen, M. B., Dalmay, T. & Baulcombe, D. C. RNA Polymerase IV Directs Silencing of Endogenous DNA. en. *Science* **308**, 118–120. ISSN: 0036-8075, 1095-9203 (Apr. 2005).
138. Wierzbicki, A. T., Ream, T. S., Haag, J. R. & Pikaard, C. S. RNA polymerase V transcription guides ARGONAUTE4 to chromatin. *Nature Genetics* **41**, 630–634. ISSN: 1061-4036 (May 2009).
139. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. en. *Science* **339**, 950–953. ISSN: 0036-8075, 1095-9203 (Feb. 2013).

140. Chao, S.-H. & Price, D. H. Flavopiridol Inactivates P-TEFb and Blocks Most RNA Polymerase II Transcription in Vivo. en. *Journal of Biological Chemistry* **276**, 31793–31799. ISSN: 0021-9258, 1083-351X (Aug. 2001).
141. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. en. *Nature* **469**, 368–373. ISSN: 0028-0836 (Jan. 2011).
142. Gilchrist, D. A. *et al.* Regulating the regulators: the pervasive effects of Pol II pausing on stimulus-responsive gene networks. en. *Genes & Development* **26**, 933–944. ISSN: 0890-9369, 1549-5477 (May 2012).
143. Min, I. M. *et al.* Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. en. *Genes & Development* **25**, 742–754. ISSN: 0890-9369, 1549-5477 (Apr. 2011).
144. Zeitlinger, J. *et al.* RNA polymerase stalling at developmental control genes in the *Drosophila melanogaster* embryo. *Nature Genetics* **39**, 1512–1516. ISSN: 1061-4036 (Dec. 2007).
145. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. en. *Science* **322**, 1845–1848. ISSN: 0036-8075, 1095-9203 (Dec. 2008).
146. Liu, W. *et al.* Brd4 and JMJD6-Associated Anti-Pause Enhancers in Regulation of Transcriptional Pause Release. *Cell* **155**, 1581–1595. ISSN: 0092-8674 (Dec. 2013).
147. Core, L. J. *et al.* Defining the Status of RNA Polymerase at Promoters. *Cell Reports* **2**, 1025–1035. ISSN: 2211-1247 (Oct. 2012).
148. Carmeliet, P. Angiogenesis in life, disease and medicine. *Nature* **438**, 932–936. ISSN: 0028-0836 (Dec. 2005).
149. Ribatti, D. & Crivellato, E. “Sprouting angiogenesis”, a reappraisal. *Developmental Biology* **372**, 157–165. ISSN: 0012-1606 (Dec. 2012).
150. Zhang, B. *et al.* A dynamic H3K27ac signature identifies VEGFA-stimulated endothelial enhancers and requires EP300 activity. en. *Genome Research* **23**, 917–927. ISSN: 1088-9051, 1549-5469 (Apr. 2013).
151. Folkman, J. Tumor Angiogenesis: Therapeutic Implications. *New England Journal of Medicine* **285**, 1182–1186. ISSN: 0028-4793 (Nov. 1971).
152. Carmeliet, P. & Jain, R. K. Molecular mechanisms and clinical applications of angiogenesis. en. *Nature* **473**, 298–307. ISSN: 0028-0836 (May 2011).
153. Cristofanilli, M., Charnsangavej, C. & Hortobagyi, G. N. Angiogenesis modulation in cancer research: novel clinical approaches. *Nature Reviews Drug Discovery* **1**, 415–426. ISSN: 1474-1776 (June 2002).
154. The CATT Research Group. Ranibizumab and Bevacizumab for Neovascular Age-Related Macular Degeneration. *New England Journal of Medicine* **364**, 1897–1908. ISSN: 0028-4793 (May 2011).

155. Los, M., Roodhart, J. M. L. & Voest, E. E. Target Practice: Lessons from Phase III Trials with Bevacizumab and Vatalanib in the Treatment of Advanced Colorectal Cancer. en. *The Oncologist* **12**, 443–450. ISSN: 1083-7159, 1549-490X (Apr. 2007).
156. Freedman, S. B. & Isner, J. M. Therapeutic Angiogenesis for Coronary Artery Disease. *Annals of Internal Medicine* **136**, 54–71. ISSN: 0003-4819 (Jan. 2002).
157. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotech* **28**, 817–825. ISSN: 1087-0156 (2010).
158. Zentner, G. E. & Henikoff, S. Regulation of nucleosome dynamics by histone modifications. en. *Nature Structural & Molecular Biology* **20**, 259–266. ISSN: 1545-9993 (Mar. 2013).
159. Schweighofer, B. *et al.* The VEGF-induced transcriptional response comprises gene clusters at the crossroad of angiogenesis and inflammation. *Thrombosis and Haemostasis* **102**, 544–554. ISSN: 0340-6245 (2009).
160. Testori, J. *et al.* The VEGF-regulated transcription factor HLX controls the expression of guidance cues and negatively regulates sprouting of endothelial cells. en. *Blood* **117**, 2735–2744. ISSN: 0006-4971, 1528-0020 (Mar. 2011).
161. Zeng, H. *et al.* Orphan nuclear receptor TR3/Nur77 regulates VEGF-induced angiogenesis through its transcriptional activity. en. *The Journal of Experimental Medicine* **203**, 719–729. ISSN: 0022-1007, 1540-9538 (Mar. 2006).
162. Ogryzko, V. V., Schiltz, R. L., Russanova, V., Howard, B. H. & Nakatani, Y. The Transcriptional Coactivators p300 and CBP Are Histone Acetyltransferases. *Cell* **87**, 953–959. ISSN: 0092-8674 (Nov. 1996).
163. Reddy, T. E. *et al.* Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Research* **22**, 860–869 (2012).
164. Bowers, E. M. *et al.* Virtual Ligand Screening of the p300/CBP Histone Acetyltransferase: Identification of a Selective Small Molecule Inhibitor. *Chemistry & Biology* **17**, 471–482. ISSN: 1074-5521 (May 2010).
165. Kundaje, A. *et al.* Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. en. *Genome Research* **22**, 1735–1747. ISSN: 1088-9051, 1549-5469 (Sept. 2012).
166. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotech* **28**, 495–501. ISSN: 1087-0156 (May 2010).
167. Boyle, A. P., Guinney, J., Crawford, G. E. & Furey, T. S. F-Seq: a feature density estimator for high-throughput sequence tags. en. *Bioinformatics* **24**, 2537–2538. ISSN: 1367-4803, 1460-2059 (Nov. 2008).
168. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing Chromosome Conformation. en. *Science* **295**, 1306–1311. ISSN: 0036-8075, 1095-9203 (Feb. 2002).
169. Randi, A., Sperone, A., Dryden, N. & Birdsey, G. Regulation of angiogenesis by ETS transcription factors. *Biochemical Society Transactions* **37**, 1248. ISSN: 0300-5127, 1470-8752 (Dec. 2009).



170. Linnemann, A. K., O'Geen, H., Keles, S., Farnham, P. J. & Bresnick, E. H. Genetic framework for GATA factor function in vascular biology. en. *Proceedings of the National Academy of Sciences* **108**, 13641–13646. ISSN: 0027-8424, 1091-6490 (Aug. 2011).
171. Hellström, M. *et al.* Dll4 signalling through Notch1 regulates formation of tip cells during angiogenesis. *Nature* **445**, 776–780. ISSN: 0028-0836 (Feb. 2007).
172. Holderfield, M. T. & Hughes, C. C. W. Crosstalk Between Vascular Endothelial Growth Factor, Notch, and Transforming Growth Factor- $\beta$  in Vascular Morphogenesis. en. *Circulation Research* **102**, 637–652. ISSN: 0009-7330, 1524-4571 (Mar. 2008).
173. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. en. *Nature Biotechnology* **31**, 46–53. ISSN: 1087-0156 (Jan. 2013).
174. Bailey, T. L. & Machanick, P. Inferring direct DNA binding from ChIP-seq. en. *Nucleic Acids Research* **40**, e128–e128. ISSN: 0305-1048, 1362-4962 (Sept. 2012).
175. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. Quantifying similarity between motifs. *Genome Biology* **8**, R24. ISSN: 14656906 (2007).
176. Robinson, J. T. *et al.* Integrative genomics viewer. en. *Nature Biotechnology* **29**, 24–26. ISSN: 1087-0156 (Jan. 2011).
177. Wythe, J. *et al.* ETS Factors Regulate Vegf-Dependent Arterial Specification. *Developmental Cell* **26**, 45–58. ISSN: 1534-5807 (July 2013).
178. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. en. *Genome Research* **22**, 1760–1774. ISSN: 1088-9051, 1549-5469 (Sept. 2012).
179. Medzhitov, R. & Horng, T. Transcriptional control of the inflammatory response. *Nature Reviews Immunology* **9**, 692–703. ISSN: 1474-1733 (Oct. 2009).
180. Kasowski, M. *et al.* Extensive Variation in Chromatin States Across Humans. en. *Science*, 1242510. ISSN: 0036-8075, 1095-9203 (Oct. 2013).
181. Kaikkonen, M. U. *et al.* Control of VEGF-A transcriptional programs by pausing and genomic compartmentalization. en. *Nucleic Acids Research*, gku1036. ISSN: 0305-1048, 1362-4962 (Oct. 2014).
182. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotech* **30**, 90–98. ISSN: 1546-1696 (2012).
183. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. en. *Nature* **489**, 83–90. ISSN: 0028-0836 (Sept. 2012).
184. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6**, 95–108. ISSN: 1471-0056 (Feb. 2005).
185. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. en. *Nature Genetics* **45**, 124–130. ISSN: 1061-4036 (2013).

186. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. en. *Nature* **518**, 337–343. ISSN: 0028-0836 (Feb. 2015).
187. Wamstad, J. A., Wang, X., Demuren, O. O. & Boyer, L. A. Distal enhancers: new insights into heart development and disease. *Trends in Cell Biology* **24**, 294–302. ISSN: 0962-8924 (May 2014).
188. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. en. *Nature* **489**, 91–100. ISSN: 0028-0836 (Sept. 2012).
189. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. en. *Genome Biology* **14**, R36. ISSN: 1465-6906 (Apr. 2013).
190. Alexa, A. & Rahnenfuhrer, J. *topGO: topGO: Enrichment analysis for Gene Ontology. R package version 2.18.0.* <<http://www.bioconductor.org/packages/release/bioc/html/topGO.html>>.
191. Storey, J., Bass, A., Dabney, A. & Robinson, D. *qvalue: Q-value estimation for false discovery rate control. R package version 1.43.0.* <<http://www.bioconductor.org/packages/release/bioc/html/qvalue.html>>.
192. Anand, P. *et al.* BET Bromodomains Mediate Transcriptional Pause Release in Heart Failure. *Cell* **154**, 569–582. ISSN: 0092-8674 (Aug. 2013).
193. Lagha, M. *et al.* Paused Pol II Coordinates Tissue Morphogenesis in the Drosophila Embryo. *Cell* **153**, 976–987. ISSN: 0092-8674 (May 2013).
194. Levine, M. Paused RNA Polymerase II as a Developmental Checkpoint. *Cell* **145**, 502–511. ISSN: 0092-8674 (May 2011).
195. Rahl, P. B. *et al.* c-Myc Regulates Transcriptional Pause Release. *Cell* **141**, 432–445. ISSN: 0092-8674 (Apr. 2010).
196. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. en. *Nature* **505**, 495–501. ISSN: 0028-0836 (Jan. 2014).
197. Gilchrist, D. A. *et al.* Pausing of RNA Polymerase II Disrupts DNA-Specified Nucleosome Organization to Enable Precise Gene Regulation. *Cell* **143**, 540–551. ISSN: 0092-8674 (Nov. 2010).
198. Hendrix, D. A., Hong, J.-W., Zeitlinger, J., Rokhsar, D. S. & Levine, M. S. Promoter elements associated with RNA Pol II stalling in the Drosophila embryo. en. *Proceedings of the National Academy of Sciences* **105**, 7762–7767. ISSN: 0027-8424, 1091-6490 (June 2008).
199. Nechaev, S. *et al.* Global Analysis of Short RNAs Reveals Widespread Promoter-Proximal Stalling and Arrest of Pol II in Drosophila. en. *Science* **327**, 335–338. ISSN: 0036-8075, 1095-9203 (Jan. 2010).
200. Boettiger, A. N. & Levine, M. Synchronous and Stochastic Patterns of Gene Activation in the Drosophila Embryo. en. *Science* **325**, 471–473. ISSN: 0036-8075, 1095-9203 (July 2009).

201. Marinov, G. K. *et al.* From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. en. *Genome Research* **24**, 496–510. ISSN: 1088-9051, 1549-5469 (Mar. 2014).
202. Yan, L. *et al.* Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. en. *Nature Structural & Molecular Biology* **20**, 1131–1139. ISSN: 1545-9993 (Sept. 2013).
203. Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. en. *Nature* **498**, 236–240. ISSN: 0028-0836 (June 2013).
204. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. en. *Nature* **503**, 290–294. ISSN: 1476-4687 (Nov. 2013).
205. Lin, C., Garruss, A., Luo, Z., Guo, F. & Shilatifard, A. The RNA Pol II Elongation Factor Ell3 Marks Enhancers in ES Cells and Primes Future Gene Activation. *Cell* **152**, 144–156. ISSN: 0092-8674 (Jan. 2013).
206. Byun, J. S. *et al.* Dynamic bookmarking of primary response genes by p300 and RNA polymerase II complexes. en. *Proceedings of the National Academy of Sciences* **106**, 19286–19291. ISSN: 0027-8424, 1091-6490 (Nov. 2009).
207. Hargreaves, D. C., Horng, T. & Medzhitov, R. Control of Inducible Gene Expression by Signal-Dependent Transcriptional Elongation. *Cell* **138**, 129–145. ISSN: 0092-8674 (July 2009).
208. Tozawa, H. *et al.* Genome-Wide Approaches Reveal Functional Interleukin-4-Inducible STAT6 Binding to the Vascular Cell Adhesion Molecule 1 Promoter. en. *Molecular and Cellular Biology* **31**, 2196–2209. ISSN: 0270-7306, 1098-5549 (June 2011).
209. Weber, C. M., Henikoff, J. G. & Henikoff, S. H2A.Z nucleosomes enriched over active genes are homotypic. en. *Nature Structural & Molecular Biology* **17**, 1500–1507. ISSN: 1545-9993 (Dec. 2010).
210. Jin, C. *et al.* H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nature Genetics* **41**, 941–945. ISSN: 1061-4036 (Aug. 2009).
211. Li, G. *et al.* Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. *Cell* **148**, 84–98. ISSN: 0092-8674 (Jan. 2012).
212. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. en. *Genes & Development* **25**, 1010–1022. ISSN: 0890-9369, 1549-5477 (May 2011).
213. Fenouil, R. *et al.* CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. en. *Genome Research* **22**, 2399–2408. ISSN: 1088-9051, 1549-5469 (Dec. 2012).
214. Henriques, T. *et al.* Stable Pausing by RNA Polymerase II Provides an Opportunity to Target and Integrate Regulatory Signals. *Molecular Cell* **52**, 517–528. ISSN: 1097-2765 (Nov. 2013).

215. Chubb, J. R. & Liverpool, T. B. Bursts and pulses: insights from single cell studies into transcriptional mechanisms. *Current Opinion in Genetics & Development. Differentiation and gene regulation* **20**, 478–484. ISSN: 0959-437X (Oct. 2010).
216. Mendenhall, E. M. *et al.* Locus-specific editing of histone modifications at endogenous enhancers. en. *Nature Biotechnology* **31**, 1133–1136. ISSN: 1087-0156 (Dec. 2013).
217. Tullai, J. W. *et al.* Immediate-Early and Delayed Primary Response Genes Are Distinct in Function and Genomic Architecture. en. *Journal of Biological Chemistry* **282**, 23981–23995. ISSN: 0021-9258, 1083-351X (Aug. 2007).
218. Michalik, K. M. *et al.* Long Noncoding RNA MALAT1 Regulates Endothelial Cell Function and Vessel Growth. en. *Circulation Research* **114**, 1389–1397. ISSN: 0009-7330, 1524-4571 (Apr. 2014).
219. Dahlman, J. E. *et al.* In vivo endothelial siRNA delivery using polymeric nanoparticles with low molecular weight. en. *Nature Nanotechnology* **9**, 648–655. ISSN: 1748-3387 (Aug. 2014).
220. Roudnicky, F. *et al.* Endocan Is Upregulated on Tumor Vessels in Invasive Bladder Cancer Where It Mediates VEGF-Induced Angiogenesis. en. *Cancer Research* **73**, 1097–1106. ISSN: 0008-5472, 1538-7445 (Feb. 2013).