A SYSTEMATIC ANALYSIS OF THE CONCORDANCE BETWEEN CHROMATIN ACCESSIBILITY AND GENE EXPRESSION CHANGES

Karun Kiani

A DISSERTATION

in

Cell and Molecular Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2022

Supervisor of Dissertation

Arjun Raj, Professor of Bioengineering and Genetics

Graduate Group Chairperson

Daniel S. Kessler, Associate Professor of Cell and Developmental Biology

Dissertation Committee

Aimee S. Payne, Professor of Dermatology
Brian C. Capell, Assistant Professor of Dermatology
Erica Korb, Assistant Professor of Genetics
Golnaz Vahedi, Associate Professor of Genetics

A SYSTEMATIC ANALYSIS OF THE CONCORDANCE BETWEEN CHROMATIN ACCESSIBILITY AND GENE EXPRESSION CHANGES

To بابا, مامی, *and Ramona,*

For giving me the kind of love that lights up the sky.

"The knowledge of anything, since all things have causes, is not acquired or complete unless it is known by its causes." -Ibn Sina, Persian polymath and father of early modern medicine

“A hair divides what is false and true.” -Omar Khayyam, Persian polymath and poet

ACKNOWLEDGEMENTS

To say it is surreal to be sitting at my desk typing these words in the early hours of the morning is a severe understatement. When I first entered this program almost six years ago, I'm not sure what I imagined reaching this point would be like or what the process would be like, but now I find myself at the terminus of this portion of the journey. This dissertation came to fruition through the help and effort of a great many individuals and reflecting on this leaves me deeply humbled and also fills my heart. On the off chance that anyone reads this, know that any omissions are not the result of some perceived slight but rather the product of a weary yet over-caffeinated mind.

First, I'd like to express my gratitude to my thesis advisor, Dr. Arjun Raj. Arjun, thank you for allowing me into your lab and sharing your wisdom over the last four years. I've learned much about science as I hope this work demonstrates, but also about how to speak, how to visualize data, and how to write. You continue to inspire me in more ways than I think you'll ever understand. Also thank you for severely and rashly underestimating the extent of my knowledge about The Lord of the Rings franchise (for the curious: the Mouth of Sauron, in fact, only appears in the extended edition of The Return of the King), allowing me to make easiest fifty dollars I will likely ever make.

To the extent an individual in their early twenties can make decisions about the next decade of their life, I must thank the individuals who inspired me to pursue the path of getting my MD and PhD. First, a great thanks is owed to Mrs. Kaye for sparking my interest in biology in high school. I also must thank my kind and wonderful undergraduate advisor Mel Coleman who not only taught and advised me, but also was kind enough to take me on adventures ranging from the cloud forests of Ecuador to the fish markets of Sapporo. To Andy Steele for taking me on in Mel's absence, for making

me giggle with his wry humor, and teaching the value of heeding the “call of the hills” (I partially blame my costly cycling habit on you). Finally, to Kumar Narayanan, for serving as an upstanding role-model as a physician-scientist and cementing my decision to pursue this program.

I am deeply indebted to the members of my thesis committee: Dr. Aimee Payne, Dr. Brian Capell, Dr. Erica Korb, and Dr. Golnaz Vahedi. It was truly a privilege to have astounding scientists of such a high caliber at my disposal for their encouragement and insights throughout my PhD. To Aimee, thank you for assuming the role of chair despite the near infinite other hats you wear throughout Penn and life, including believing me enough early on to grant me acceptance into the wonderful Penn MSTP community. To Brian, thank you for helping me explore my career options as a clinician and welcoming me into your clinic (career decisions TBD). To Erica, thank you for your extremely astute questions during my committee meetings to the point that I always got a little scared when you chimed in (but they always raised insightful and important points in a kind and thoughtful way) and for somehow making that one committee meeting on time despite a flat tire. And to Golnaz, a special thanks for welcoming me into your lab for a rotation where many of the skills in analyzing ATAC-seq data vital to this dissertation were developed and for modeling Persian excellence.

I also want to thank the many people with foundational roles in the working of the programs that make this thesis possible. Many thanks to Dr. Skip Brass for believing that I too, could join the greatest MSTP in the galaxy™. A special thank you to Maggie Krall, the person without which the sun would not rise every day, for your continuing support and counsel. I could write songs about how much you do for everyone in the MSTP. I also would like to thank the various people involved in CAMB, G&E, and Penn

Genetics, especially Kimberly Runyan, Meagan Schofer, and Christina Streathearn. Thank you to my funding source, NIH T32 GM008216 and Dr. Doug Epstein for your mentorship and vote of confidence in me. A special thank you to Dr. Christina Leslie and Alvaro González who were so incredibly patient and receptive to my many questions regarding their work and who this dissertation would not be possible without.

Next, I must acknowledge all former and current members of the Raj lab, for being exemplary colleagues and creating a wonderful work environment to learn and train in. To Ryan Boe and Chris Coté for being able to discuss the most obscure corners of pop culture. To Maggie Dunagin for all you do for the lab to keep the ship afloat. To Drs. Ally Coté and Aman Kaur for your guidance and collaboration. To Sam Reffsin for always being good for not drinking coffee in the lounge alone with. To Dr. Ben Emert for patience in teaching me wet lab techniques. To Dr. Lee Richman for his mage-like knowledge of R and being great company for a bouldering sesh. To Dr. Eduardo Torre for being the best company for happy hour. To Dr. Ian Mellis for continuing to inspire as a scientist and as a person, and who can make a hell of a cold yaki soba noodles. Finally, to Dr. Lauren Beck, for sharing my enthusiasm for hip-hop, herbal Italian liquors, and Euphoria, and for making me the five song country playlist.

I am extremely fortunate to have a rich life filled with friends outside the lab who I have leaned on extensively throughout this time. To my friends from my time at Claremont McKenna: Maddie Bannon, Carly Goodkin, Abby Dolmseth, Gracie Mahan, Megan Latta, Elica Sharifnia, Ben Goldberg, Danial Ceasar, Jackson Badger, the entire Duckworth family, Will Sippl, Maya Reddy, India Wade, and Kelsey Gross. Thank you for being my friend during my more formative years and not immediately running in the opposite direction and it has been truly a treat to keep growing and going through life

with you all. To my friends from childhood and my time living in Boston: John and Joseph Replogle, Michael Baranowski, Billy McCormick, Garrett Wong, and Yoshi Rothman. To my friends I've made during medical school and the MSTP: Scott Symmonds, Amy Campbell, Elle Lett, Aaron Williams, Eli Cornblath, Angela Song, Bob Lou, Ming Cai, Natalie Gong, Bridget Gosis, Nate Sanford, Paul & Marquise, Nikki Johnson, Emily Rider-Longmaid, Sam Huo, Josh Ho, and Lauren Reed-Guy, your friendship has lifted my spirits and inspired me throughout this long journey.

To my close friends from my Penn MSTP cohort. To Selen Uman and Fola Sofela, my darling anatomy team partners, no better company can be found whether dissecting a cadaver or drinking awful dollar cocktails. To Juan Serrano, for always welcoming us into your home and tolerating my awful Halo skills. To Steph Teeple for being extremely cool and sharing a love of pretty bad movies. To my dear, dear, dear friends and buds John Bernabei, Daniel Park, Daniel Xu, what a privilege to know you and grow along with you all these last few years, I truly would never make it to this point without you. I hope whenever you feel down you think of the magic of that night on the Kamogawa.

Next comes a category of individuals who in truth deserve to have entire chapters dedicated to them for their contributions, but I will try to be brief. To my cousin Tina for being one of my oldest friends and my dear uncle Darius for always making me laugh. To Bug, my closest friend for the better part of a decade through multiple breakups, a cancer diagnosis, trips around the globe, and cohabitation for four of my most excellent years, thank you for inspiring me everyday. To Kate, a friend who I never have to hide my authentic self from, who understands my fundamental need to speak like Jar Jar Binks, you are truly the best of us. To Eric, my roommate, my collaborator, and kindred spirit. I wish I could put into words the deep joy I feel when I swing the door open and see your

profile seated on the couch. You will both be sorely missed as you move on to the next stage, but don't for a second think you'll ever get rid of me. To my dear Connie, my first and fast friend in the Raj Lab. For entertaining my rants on topics esoteric, for making sure I am fed when sick, for helping me channel into my creative side, for teaching me to actually use illustrator, and modeling being a physician-scientist among so many other things. To Yogesh, the David Moyes to my Jesse Lingard (though current events have rendered this analogy null), for teaching me more than possibly any other individual has. For leaving me inspired after every conversation, and for whom I would not be writing these words without your help. For keeping me company in lamenting the mediocrity of our respective football clubs. Finally, to Naveen, the Pippin to my Merry. For being my friend, for understanding me so fundamentally that words often do even need to be said. For making every day of the PhD filled with a ray of sunshine regardless of anything else. For always searching for the sublime and inspiring and pushing me to be my best self. To all of you I have no other words other than thank you from the deepest depths at the bottom of my heart.

To my partner and my family, a tremendous thanks is owed for your love. Thank you to Virginia for making me smile, making me laugh, making me coffee, and accompanying me through life. Words fail me, but it is my sincere hope that you can feel even a small portion of the extent of gratitude I have for knowing you. To my brother-in-law Danny for continued support and serving as a much-needed foil at times to the rest of my family. To my oldest friend, role model, and big sister Ramona. I am proud of our relationship and can feel your love and support even from thousands of miles away. To my nephew Cameron, who is already one of my favorite people, thank you for reminding me of the importance of joy. Finally, to my beloved parents, Maryam Jafari and Mansour

Kiani, the spring from which the river flows. I do not know if I can ever fully express my gratitude or honor the sacrifices you have made for all you have given me in this life. For making me the person I am, for inspiring me, it is my most heartfelt hope that you understand my deep appreciation. Thank you will never be enough.

ABSTRACT

A SYSTEMATIC ANALYSIS OF THE CONCORDANCE BETWEEN CHROMATIN ACCESSIBILITY AND GENE EXPRESSION CHANGES

Karun Kiani

Arjun Raj

A major goal in the field of transcriptional regulation is the mapping of changes in the binding of transcription factors to the resultant changes in gene expression. Recently, methods for measuring chromatin accessibility have enabled us to measure changes in accessibility across the genome, which are thought to correspond to transcription factor binding events. In concert with RNA-sequencing, these data in principle enable such mappings; however, few studies have looked at their concordance over short duration treatments with specific perturbations. Here, we used tandem, bulk ATAC-seq and RNA-seq measurements from MCF-7 breast carcinoma cells to systematically evaluate the concordance between changes in accessibility and changes in expression in response to retinoic acid and TGF-β. We found two classes of genes whose expression showed a significant change: those that showed some change in accessibility of nearby chromatin, and those that showed virtually no change despite strong changes in expression. The peaks associated with genes in the former group had a lower baseline accessibility prior to exposure to signal. Analysis of paired chromatin accessibility and gene expression data from distinct paths along the hematopoietic differentiation trajectory showed a much stronger correspondence, suggesting that the multifactorial biological processes associated with differentiation may lead to changes in chromatin accessibility that reflect rather than drive altered transcriptional status. Together, these results show many gene

expression changes can happen independent of changes in accessibility of local chromatin in the context of a single-factor perturbation and suggest that some changes to accessibility changes may occur after changes to expression, rather than before.

Furthermore, we establish the role of cell-intrinsic differences in clonal melanoma cell lines leading to a rare subpopulation of cells that demonstrate invasive behavior both *in vitro* and *in vivo*. This population is molecularly characterized by the high expression of *SEMA3C*, and knockout studies demonstrate that the formation of the invasives subpopulation is negatively regulated by the transcription factor *NKX2.2*. Overall, these results establish a role for non-genetic differences in important cancer attributes such as cellular invasion.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

I have always been inspired by the natural world. For example, *Agave americana* is a flowering plant referred to more colloquially as the century plant that is native to the semi-arid to arid climates of Northern Mexico and the American southwest. The century plant, despite its moniker, lives only for about 10-30 years and has an interesting strategy for reproduction: at the end of its life cycle, it sprouts a flowering stalk that can stand up to 8 m tall to spread seeds and then soon thereafter dies. Every time I notice a century plant in bloom I cannot help but wonder: how is it that one year the plant may continue business as usual, while eventually, when the conditions are opportune, the plant activates a complex genetic program to massively rewire its metabolism and reroute all of its energy into growing and maintaining the flowering stalk as long as possible ("Agave Americana (century Plant)" n.d.)?

To borrow concepts from control theory, I have been fascinated by how dynamical biological systems are constantly able to maintain or achieve new states of stability through the interaction of controllers, systems, and sensors (Strogatz et al. 1994; Åström and Murray 2010). Specifically, in examples like the above, sprouting in the century plant or other biological processes ranging from glucose homeostasis to transdifferentiation to development, sensors constantly probe the environment and activate or suppress complex gene regulatory programs in response to what may or may not be going on.

The advent of recent technological advances in sequencing methods as well as computational approaches to analyze these data have allowed for the quantification of genetic, epigenetic, and transcriptional states of populations of cells as well as individual

cells at an unprecedented level. This abundance of data provides an opportunity to use the principles of systems biology to begin to understand the interplay through which transcriptional regulation is achieved.

The transcription of genes represents an output that results from the interplay of nucleotide sequence, transcription factor binding, chromatin accessibility, and DNA methylation, among a host of other factors. Throughout this dissertation my hope was to pop the metaphorical hood of transcriptional regulation in dynamical biological systems and begin to better understand and contribute to the conversation by examining the relationship between different modalities of sequencing data (**Chapter 2**), and better understanding the heterogeneity of phenotypes relevant to the natural history of disease (**Chapter 3**).

1.1 Transcription factors and control of gene expression

A complex interplay of *cis*-regulatory elements such as promoters, enhancers, silencers, and insulators, as well as transcription factors, control the level of gene expression, which is important for development, establishing cell identity, and coordinating transcriptional responses to a variety of stimuli. Genome-wide studies have estimated 200-300 transcription factors that bind directly to core promoter elements while an additional approximately 1400 transcription factors that bind to a specific DNA sequence and thus regulate only a specific subset of genes.

The importance of proper transcription factor control of gene expression is underscored by the large corpus of evidence demonstrating the association of transcription factor dysfunction with a wide array of human diseases ranging from cancer to development disorders (Jimenez-Sanchez, Childs, and Valle 2001).

1.1.1 Structure and syntax of transcription factor activity

Transcription factors typically recognize degenerate DNA sequences that are 6-12 base pairs in length. This straightforward sequence specificity suggests that additional, more complex, rules are involved in controlling transcriptional output. One such level of control is requiring combinatorial input of transcription factor binding. Transcription factors often bind to enhancer regions that contain focused clusters of transcription factor binding sites such that combinatorial binding can result in precise and distinct patterns (Spitz and Furlong 2012). For example, in the fruit fly *D. melanogaster,* the pMAD, the phosphorylated form of the transcription factor MAD, provides part of the ability for cells to adopt a particular fate during development. It is the combinatorial binding with cell-type specific transcription factors that confers cell-fate specification, such as Tinman (Xu et al. 1998; H.-H. Lee and Frasch 2005) in the dorsal mesoderm or Scalloped (Guss et al. 2001) for the wing imaginal disc.

Another method of control of transcription factor activity comes from not only the timing of transcription factor expression, but also the timing of its DNA-binding activity as demonstrated for mammalian myoblast differentiation with MYOD1 where some enhancers are continuously bound while others are bound at only early or late stages in development (Cao et al. 2010).

Finally, cooperativity serves to modulate transcription factor activity. This occurs with regards to both (1) indirect effects that do not affect affinity for the cognate motif and (2) direct effects that alter motif affinity. In the former case, the activity of two transcription factors binding at a given enhancer can in some cases lead to increased occupancy for each transcription factor (Voss et al. 2011). This occurs through both synergistic action between the two transcription factors and nucleosome repositioning,

termed 'assisted loading,' as well as local bending of DNA which assists binding. In the latter case, binding to a protein cofactor can affect the affinity of the two factors for their respective motifs (Spitz and Furlong 2012). More recent works have also demonstrated an alternative mechanism where protein-protein interactions can altogether change a transcription factor's DNA sequence specificity. For example, homeodomain-containing Hox proteins in *D. melanogaster* when interacting with extradenticle (EXD) proteins can induce subtle changes to DNA-binding specificities (Slattery et al. 2011). Thus, transcription factors use a variety of methods to interact with local chromatin in order to facilitate changes in gene expression.

1.1.2 Measuring the binding of transcription factors to DNA

Technologies that make use of chromatin immunoprecipitation combined with microarray (ChIP-chip) or high-throughput sequencing (ChIP-seq) have been instrumental in determining transcription factor-DNA binding patterns genome-wide and still serve as the "gold-standard" for measuring protein-DNA interactions. Briefly, they crosslink DNA-protein interactions and employ the use of a specific antibody targeting the transcription factor of interest to pull down and subsequently sequence only the chromatin sequences that are bound to the transcription factor. Similarly, transcription factor-DNA interactions can be mapped globally using DamID, which uses a fusion of a DNA methyltransferase domain to the DNA-binding protein of interest to identify regions with adenine methylation. Other chromatin accessibility measuring methods (mentioned later in this introduction) can also indirectly infer the activity of transcription factors by looking for their "footprints" at sites of accessible chromatin.

One can categorize the binding sites from these methods based on features such as nearest gene, the relative frequency of regions relative to gene structure (e.g. promoter,

intron, exon, etc.), or the type of chromatin domain. These data can then be compared between conditions, cell lines and cell types, or transcription factors to begin to understand the underlying regulation and logic of transcription factor binding.

1.1.3 Integration of ChIP-seq data with genome-wide expression data from RNA-seq

Combining whole-genome transcriptomic data from RNA-sequencing (RNA-seq) with transcription factor binding profiles from ChIP-seq provides a valuable opportunity to study the interplay between transcription factor binding and gene transcription with implications for both normal physiology and disease pathogenesis (Feng et al. 2014). While there exist many tools that can be used "out of the box" for the analysis of these data independently (Angelini and Costa 2014), the integration of these data types is not trivial and presents one of the greatest challenges in modern biology (Gomez-Cabrero et al. 2014). Early approaches use a variety of frameworks including a log-linear regression model (Ouyang, Zhou, and Wong 2009) or a support vector regression (Cheng, Yan, Yip, et al. 2011; Cheng, Yan, Hwang, et al. 2011; Cheng et al. 2012; Dong et al. 2012; Cheng and Gerstein 2012). Many of these approaches attempted at using transcription factor binding or histone modification data to predict gene expression within a condition. Using these data to correlate variations of features of epigenetic marks and gene expression between two conditions adds another level of complexity, and usually these methods only show predictive power when categorizing gene expression as a categorical output variable (i.e. upregulated, downregulated, or unchanged expression) (Klein et al. 2014). However, an important caveat of these approaches is that ChIP-seq data for transcription factor binding is mainly based on a given binding peak's proximity to the transcriptional start site, and many existing approaches rely on local interaction. Owing

to the paucity of data and extra analysis considerations of techniques that map long-range chromatin interactions such as HiChIP and Hi-C, their integration with transcriptomic data has been relatively under-explored (Angelini and Costa 2014).

1.2 Chromatin accessibility and gene regulation

The core structural unit of DNA packaging is referred to as a *nucleosome*, which consists of approximately 147 base pairs of DNA wrapped around a hetero-octamer of positively charged histone molecules much like a garden hose is wrapped around a reel. The nucleosome serves as the cornerstone for the multiple layers of topological complexity that allows the almost 3 meters of DNA to be confined within the volume of the nucleus of eukaryotic cells.

Nucleosomal DNA is not evenly distributed across the genome, and varies greatly between cell types, and even within the same cell type depending on context. DNA bound to histone molecules is referred to as relatively inaccessible while nucleosome-free DNA is thought to be more accessible in that it can be bound by other DNA-interacting macromolecules such as transcription factors, architectural proteins, or polymerases. It is helpful to conceptualize DNA along a continuum of inaccessible to accessible because it has widespread consequences for gene regulation. For example, nucleosome-depleted regions are commonly thought to represent non-coding DNA regions that are involved in the regulation of expression of nearby genes, termed *cis*-regulatory elements and include enhancers and promoters. *Cis*-regulatory elements exert these effects by interacting with transcriptional regulators such as transcription factors. Indeed, while the accessible portion of the genome is only approximately 3% of the total genome, over 90% of transcription factor binding sites are confined to this accessible compartment (Thurman et al. 2012). Nucleosomes are dynamic in terms of their positioning along the genome,

their assembly and disassembly, and the myriad of post transcriptional modifications. Thus, the topology of chromatin within 3 dimensional space provides an important point of regulation of transcription.

1.2.1 Measuring bulk chromatin accessibility

Methods for measuring chromatin accessibility are based on enzymes being able to physically interact with accessible chromatin in order to fragment, tagment, or chemically label (e.g., methylate) these accessible portions of DNA (Boyle et al. 2008; Schones et al. 2008; Hesselberth et al. 2009; Kelly et al. 2012; Buenrostro et al. 2013; Minnoye et al. 2021). Early experiments in the 1970s used the endonuclease deoxyribonuclease I (DNase I) to show promoters and introns of expressed genes are more sensitive than other regions to digestion by DNase I, indicating that the chromatin is particularly accessible in these regions.

With the establishment of high-throughput sequencing technologies, these enzymatic methods could be combined with sequencing to begin to resolve chromatin accessibility genome-wide. For example, DNase I hypersensitive site sequencing (DNase-seq) was one of the first instantiations of this approach and still is the approach of choice for transcription factor footprinting, which can identify the location of transcription factor binding sites due to the protection of the local chromatin from the transcription factor itself (Hesselberth et al. 2009).

Using a micrococcal nuclease (MNase) in a technique called MNase-seq leverages the ability of MNase to both act as an endonuclease to cleave internucleosomal DNA and an exonuclease to degrade DNA not protected by proteins. This ability makes MNase-seq particularly useful for isolating DNA fragments spanning a single nucleosome (West et al. 2014). Alternatively, nucleosome occupancy and methylome sequencing (NOMe-seq)

chemically modifies rather than cleaves accessible DNA using a GpC methyltransferase to create sites of ectopic methylation of CG dinucleotides (nb: the endogenous methylation of DNA found in both the human and mouse genomes occurs at *CG* dinucleotides) (Kelly et al. 2012). NOMe-seq does not rely on any enrichment-based steps and therefore requires a greater read depth compared to other methods. However, this necessity also proves to be an advantage because it creates a more quantitative measurement of accessibility compared to other techniques (Kelly et al. 2012; Minnoye et al. 2021).

However, the current genome-wide chromatin accessibility method *du jour* for almost the last decade has been Assay for Transposase-Accessible Chromatin and sequencing, or ATAC-seq (Buenrostro et al. 2013) or one of its derivative variants (Corces et al. 2016, 2017). A genetically engineered hyperactive transposase (Tn5) is preloaded with Illumina adapters to simultaneously cleave and tag accessible chromatin regions. These target DNA fragments are purified and amplified via PCR before being sequenced using high-throughput technologies (Buenrostro et al. 2013). The major advantages of ATAC-seq and its variants in that they require relatively low sample input (500-50,000 cells versus millions needed for DNase-seq) and the protocol takes less than a day to completely compared to the few days needed for DNase-seq or MNase-seq (Buenrostro et al. 2013; Minnoye et al. 2021).

1.2.2 Chromatin accessibility at single-cell resolution

Advances in barcoding and microfluidic technologies have allowed the measurement of chromatin accessibility at single-cell resolution. While these exist for many of the methods mentioned for bulk profiling (e.g. scDNase-seq (Jin et al. 2015) or scMNase-seq (Lai et al. 2018)), single-cell ATAC-seq (scATAC-seq) has become a popular approach

due to its relative simplicity and reproducibility (Buenrostro et al. 2015; X. Chen et al. 2018; Cusanovich et al. 2015; Lareau et al. 2019; Satpathy et al. 2019).

Droplet based methods exist in multiple commercially available kits (e.g. Chromium Next Gem Single Cell ATAC-seq or SureCell ATAC-seq) that when combined with standard sequencing library reagents and proprietary robotic sample processing devices allow for a relatively straightforward and reproducible approach to scATAC-seq. Alternatively, plate-based methods (X. Chen et al. 2018; Mezger et al. 2018) require single cells to be sequestered into individual wells of a plate but this approach limits overall throughput of the assay. Given limitations and relative novelty of these technologies, they will, for the most part, not be the focus of this dissertation but their future use is discussed in more detail in **Section 5.2** of chapter 5.

1.2.3 Integration of ATAC-seq and RNA-seq data

Many studies have examined the relationship between chromatin accessibility data and gene expression data to some degree or another in an attempt to better understand underlying regulation (González, Setty, and Leslie 2015; Ackermann et al. 2016; Ampuja et al. 2017; Ramirez et al. 2017; de la Torre-Ubieta et al. 2018; Starks et al. 2019; Bunina et al. 2020; Hota et al. 2022), but the analyses usually are the subject of at best one panel of one figure. They often either focus on accessibility measurements at or near the promoter only (Ampuja et al. 2017) or look at only at the relationship between chromatin accessibility and gene expression (Starks et al. 2019) rather than the relationship between their *change* over time or in response to some perturbation. Here, we sought to rigorously and systematically characterize the concordance, or lack thereof, between chromatin accessibility and gene expression data in response to single-factor perturbations (Sanford et al. 2020; Kiani et al. 2022) to better understand the

underlying logic of transcription in this context as well as limitation of the technologies involved.

1.3 Cell-intrinsic differences and metastasis in melanoma

1.3.1 Disease and metastasis background

Owing to the high metastatic potential of cutaneous melanoma, it has been estimated by the SEER database that the 5-year overall survival rate of stage IV metastatic melanoma is approximately 30%, a value that has scarcely changed in the last twenty years (Song et al. 2015). Metastasis refers to the result of a complex series of events where cancer cells are able to leave the primary tumor and travel via lymphatic or hematological spread before arriving at a different anatomical site and resuming proliferation. These events are characterized by molecular changes that lead to distinct cellular phenotypes. In particular, a cell must first become invasive to leave the primary tumor and migrate to the site of metastasis and subsequently re-adopt a proliferative phenotype to establish the metastatic nidus (Polyak and Weinberg 2009; Mittal 2018). Moreover, typically very few cells from the original tumor will undergo the necessary steps to metastasize (Francí et al. 2006; Mani et al. 2008).

1.3.2 Cell-autonomous differences and metastatic potential

While there exists evidence to support that the rare cells that do in fact leave the primary tumor to metastasize do so due to external factors such as the tumor microenvironment (Olmeda et al. 2017; Kaur et al. 2019), less work has characterized whether or not cell-intrinsic factors prime certain cells for metastasis. More classically, the erstwhile factors are considered to be mutations that drive an increased capacity for metastasis (Nataraj, Marrocco, and Yarden 2021; Nguyen et al. 2022). However, more recently, the role of

non-genetic changes in transcriptional regulation have also been implicated in driving the phenotype switching to an invasive identity (Arozarena and Wellbrock 2019; Quinn et al. 2021). Indeed, there is precedent demonstrating that non-genetic differences lead to distinct behaviors in biology and cancer, specifically in the context of therapy resistance (Symmons and Raj 2016; A. Raj and van Oudenaarden 2008; Emert et al. 2021; Goyal et al. 2021; E. A. Torre et al. 2021; Shaffer et al. 2017; Sharma et al. 2010; Gupta et al. 2011). What is far less explored is how these non-genetic differences may contribute to invasiveness in melanoma. Changes in expression of genes such as *ALDH1A1*, *KIT*, and *HSP90AB1* have been implicated in distinguishing metastatic melanomas from their primary tumor counterparts (Metri et al. 2017; Turner, Ware, and Bosenberg 2018).

1.4 Summary

Through two distinct, but related themes, in this dissertation I seek to better gene expression regulation both through the context genome-wide omics methods to developing new technologies to interrogate the role and consequences of non-genetic heterogeneity to using established bioinformatic methods to better understand rare invasive behavior in melanoma. As a whole, these themes move the field forward by creating a framework for systematically examining the interplay of genome-wide chromatin accessibility and gene expression data, allowing the combination of single-cell transcriptomics with lineage information, and establishing a basis for a novel role of *NKX2.2* in melanoma metastatic potential.

CHAPTER 2: CHANGES IN CHROMATIN ACCESSIBILITY ARE NOT CONCORDANT WITH TRANSCRIPTIONAL CHANGES FOR SINGLE-FACTOR PERTURBATIONS

2.1 Introduction

2.1.1 Transcription factor activity and changes in gene expression

Transcription factors regulate gene expression by binding to specific DNA sequences, facilitating transcription through the recruitment and activation of the transcriptional machinery. Deciphering the combinatorial logic underlying which transcription factors bind to what portions of DNA and in what contexts is a central challenge in creating a complete model of transcriptional regulation. Sequencing-based methods have enabled the measurement of transcript levels for all genes as well as the putative binding profiles of transcription factors across the genome. However, the precise mapping between changes in these putative binding profiles and the changes in transcriptional activity remain the subject of debate.

2.1.2 Measuring transcription factor activity

A key component of decoding the relationship between transcription factor activity and the resultant changes in transcription is the measurement of transcription factor binding to DNA. Recently, the combination of biochemical binding assays with sequencing-based readouts has led to a cornucopia of methods for making such measurements. One workhorse method is chromatin immunoprecipitation sequencing (ChIP-seq), which characterizes the binding of transcription factors and other DNA-protein interactions genome-wide (Barski et al. 2007; Robertson et al. 2007; Ma and Zhang 2020) by using

immunoprecipitation of proteins that bind to chromatin and subsequently sequencing the coprecipitated DNA. However, ChIP-seq is limited in that each experiment can only interrogate the binding profile of one transcription factor at a time.

An alternative approach that circumvents that issue is the measurement of changes in accessibility of DNA to infer changes in the binding of all transcription factors at once. Accessible regions of DNA (i.e. those regions depleted of nucleosomes) represent only 3% of the genome, but often participate in the regulation of gene expression (Weintraub and Groudine 1976; C. Wu, Wong, and Elgin 1979; C.-K. Lee et al. 2004; Thurman et al. 2012). These regions can be detected genome-wide by combining the enzymatic activity of nucleases with high-throughput sequencing using techniques such as DNase I hypersensitive site sequencing (DNase-seq) (Boyle et al. 2008) and assay for transposase accessible chromatin with sequencing (ATAC-seq) (Buenrostro et al. 2013). The interpretation of these accessibility methods leans heavily on the assumption that changes in regulatory factor binding are reflected in changes in chromatin accessibility. Certainly, there are many examples in which the correspondence between changes in accessibility strongly correspond to changes in transcriptional output. For instance, summation of ChIP-seq signal for 42 transcription factors mapped by encode in K562 chronic myelogenous leukemia cells paralleled the signal from accessible sites revealed by DNase-seq (Thurman et al. 2012). Moreover, computational methods to infer transcription factor footprints from accessibility measurements have been shown to recapitulate ChIP-seq binding well (Pique-Regi et al. 2011). Accessibility methods can also be used to look for changes in accessibility across various perturbations and cell types. Changes in accessibility generally seem to correspond to changes in transcription in the sense that large changes in transcriptional output are reflected in broad changes in

the accessibility of several loci in the surrounding chromatin (González, Setty, and Leslie 2015; de la Torre-Ubieta et al. 2018).

2.1.3 The relationship between chromatin accessibility and gene expression changes

However, it is unclear how well these accessibility based methods capture the activity of all transcription factors. It is possible that some transcription factors' binding and activity does not result in corresponding changes in accessibility and vice versa. Such a lack of correspondence could manifest itself as a lack of correlation between changes in accessibility and changes in transcription. Given the underlying assumption that a change in transcription must be mediated by the change in some transcription factor activity, then such a lack of correspondence would suggest that changes in the activity of transcription factors could change expression without changing accessibility near its binding site. While reports from the literature generally show a strong correspondence (de la Torre-Ubieta et al. 2018; González, Setty, and Leslie 2015; Ampuja et al. 2017; Starks et al. 2019), it is worth noting that the comparisons in such studies are often across rather different cell types. In such cases, it is possible that the changes in accessibility are not driven by regulation *per se*, but rather reflect the consequences of sequential exposure to multiple regulatory factors that characterize the differentiation process. Such accessibility changes could, in principle, signify the reinforcement of genes that are already transcriptionally active genes, or could even just appear around actively transcribed genes without any functional role. Disentangling such possibilities could be revealed with the use of single-factor perturbations that more directly affect an individual pathway; however, few such data are available.

Here, we used tandem bulk RNA-seq and ATAC-seq data from MCF-7 breast carcinoma cells exposed to multiple doses of retinoic acid or TGF-β to determine the degree of concordance between changes in chromatin accessibility and changes in gene expression. Furthermore, we evaluated concordance in another published data set of hematopoietic differentiation to validate our approach based on well-defined and specific perturbations. We demonstrate that while some differentially expressed genes have a high concordance between gene expression and chromatin accessibility changes, many other genes are differentially expressed without changes in their local chromatin accessibility.

2.2 Results

2.2.1 Genome-wide expression and chromatin accessibility changes reflect known biology of two perturbations

To measure the correspondence between changes in chromatin accessibility and changes in gene expression, we used MCF-7 breast carcinoma cells due to their previously described transcriptional responses to all-trans retinoic acid (Hua et al, 2009) (referred to from here on as retinoic acid) and transforming growth factor beta (TGF-β) (Mahdi et al, 2015). We used paired, bulk accessibility (ATAC-seq) and expression data (RNA-seq) from these cells (Sanford et al, 2020) collected 72 hours after continuous exposure to three different doses of each signal (**Figure 2.1**). We chose this timescale because previous work with MCF-7 cells showed more transcriptional changes at 72 hours compared to 24 hours after exposure to retinoic acid (Hua et al, 2009), and chromatin accessibility changes may not be detectable until 24 hours after perturbation (Ramirez et al, 2017). Differential gene expression and differential peak accessibility analysis showed

a dose-dependent response to both signals compared to ethanol control (**Figure 2.1**, bar plots). The ethanol 'vehicle' controls comprise three different densities of cells, and the transcriptomes of control conditions globally were similar regardless of cell density (**Figure 2.2**). To confirm that global gene expression and chromatin accessibility patterns were similar between signals and dosages, we performed principal component analysis. For both RNA-seq and ATAC-seq data, all samples exposed to the same signal or ethanol control clustered together, indicating that their gene expression and chromatin accessibility were more similar to each other than to other conditions, supporting the quality of these data.

To validate that changes in gene expression were consistent with the known biology of these signaling pathways, we performed over-representation analysis on the upregulated genes in response to high dose retinoic acid or TGF-β against curated gene sets from the molecular signatures database (Liberzon et al, 2011, 2015). The top ten gene sets based on false discovery rate (FDR)-adjusted p-values were processes canonically associated with retinoic acid (morphogenesis, organ development, anterior-posterior patterning) and TGF-β (extracellular matrix, endopeptidase activity), respectively (**Figure 2.3**). Gene set enrichment analysis (Subramanian et al, 2005) showed that genes that were differentially expressed in response to high dose retinoic acid were significantly enriched for genes associated with skeletal system morphogenesis, and genes that were differentially expressed as a result of exposure to high dose TGF-β were significantly enriched for genes associated with epithelial-to-mesenchymal transition (**Figure 2.4**). Thus, the differentially expressed genes generally reflected the known biology of the signals the cells were exposed to.

We next wondered if the changes in chromatin accessibility in response to signal were associated with the activity of specific transcription factors, in particular, those associated with the biology of these signaling pathways. We used a modified version of the chromVAR package along with its curated database of transcription factor motifs, cisBP, to identify the transcription factors with the largest predicted change in activity (Schep et al, 2017). We used the set of differential peaks to determine the set of the top 150 transcription factors with the greatest magnitude of change. These included the binding motifs of transcription factors that are canonical effectors of retinoic acid (RAR-α, HOXA13) and TGF-β signaling (SMAD3, SMAD4, and SMAD9). For each of these transcription factor motifs, we calculated a motif enrichment score for each condition based on the bias-uncorrected deviation score from chromVAR. The motif enrichment score represents the percentage change in ATAC-seq fragment counts in all peaks that contain a given transcription factor's motif (**Figure 2.5**). For example, the enrichment score of 28% for SMAD3 in the TGF-β condition meant that peaks containing the SMAD3 motif on average saw a 28% increase in fragment counts after exposure to TGF-β. We pooled together the low, medium, and high doses for each condition together in order to decrease the variability of motif enrichment scores estimates. Thus, our data recapitulated expected changes in accessibility, presumably due to the activity of transcription factors well-known to be activated by the signals used. Thus, of the changes in accessibility we did detect, they made sense based on a model of transcription factor activity leading to changes in accessibility. However, it was still possible that the activity of many transcription factors was not captured by changes in accessibility.

2.2.2 The relationship between changes in chromatin accessibility and gene expression varies on a gene by gene basis

We next wondered whether genes that were differentially expressed were more likely to have differentially accessible peaks nearby, i.e., was there concordance between gene expression and chromatin accessibility changes at the level of individual genes? To characterize the extent of concordance between these data, we looked at the overlap between genes that were differentially expressed in response to high dose signal and genes with differentially accessible peaks nearby after exposure to signal (**Figure 2.6**). We assigned each accessible peak to the nearest transcriptional start site ("nearest approach") and found that of the over 2000 genes upregulated in response to high dose retinoic acid, more than half of them had at least one differential peak assigned to its transcriptional start site (p-value < 2.2x10-16, Fisher's exact test). Similarly, a third of the genes whose expression was upregulated in response to TGF-β had differential peaks assigned to them (p-value < 2.2x10-16, Fisher's exact test). Thus, genes that are differentially expressed are more likely than random chance to have a nearby peak that is differentially accessible in response to retinoic acid or TGF-β.

While using this overlap-based approach showed correspondence between genes that are differentially expressed and their nearby peaks in response to signal, aspects of the nature of the concordance of these changes were not captured by this analysis. For example, the overlap-based method counted all differentially accessible genes that had at least one differentially accessible peak assigned to them as concordant, but did not take into account the proportion or degree to which those nearby peaks change. Moreover, we did not take into account the relationship between directionality of changes in gene expression and chromatin accessibility. The underlying assumption at the basis of this

relationship is that when peaks become more accessible that the nearby gene increases its expression, and the overlap-based approach does not take this correspondence of the direction of change into account. To better characterize these facets of concordance, we first individually examined the changes in chromatin accessibility nearby two genes whose expression were upregulated in response to retinoic acid.

HOXA1 and SLC5A5 induction are associated with exposure to retinoic acid (Glover et al, 2006; Schmutzler et al, 1997; Kogai et al, 2000), and both genes showed a dose-dependent increase in expression in response to retinoic acid (**Figures 2.7A, B**). After optimizing parameters for calling peaks and determining differentially accessible peaks (**Figure 2.8**), we found that while a large number of peaks are differentially accessible near the HOXA1 locus (**Figure 2.7A**, track view middle, black traces in accessibility plot, right), very few peaks are differentially accessible near the SLC5A5 locus (**Figure 2.7B**, track view middle, accessibility plot, right). Therefore, genes with high expression change in response to signal can show a large degree of accessibility changes or show very little accessibility changes, suggesting that changes in transcription factor activity may or may not be reflected in changes in accessibility.

2.2.3 Chromatin accessibility changes are less concordant with large changes in gene expression in signaling compared to hematopoietic differentiation

Next, we evaluated the concordance between accessibility and gene expression genome-wide while also factoring in the directionality of changes and the relative proportion of peaks that are changing on a gene-by-gene basis. As a point of comparison, we used previously published gene expression and chromatin accessibility data from hematopoietic differentiation (González et al, 2015) that demonstrated that large

changes in gene expression were typically associated with gains or losses (depending on the direction of expression change) of cell type-specific enhancers when comparing the expression and accessibility of hematopoietic stem and progenitor cells (HSPCs) to monocytes.

Before using this data set as a comparison to ours for measuring concordance between chromatin accessibility and gene expression changes, we verified that the hematopoietic differentiation data was similar to our own by a variety of metrics. First, we wanted to compare whether the number of differentially expressed genes and differentially accessible peaks between HSPCs and monocytes in the hematopoietic differentiation data was similar to the numbers from MCF-7 cells exposed to retinoic acid or TGF-β. We found that both HSPC and monocyte populations had greater than 2000 genes that were specifically expressed in their respective cell types compared to the approximately 2000 and 1500 genes differentially expressed in MCF-7 cells in response to high dose retinoic acid and TGF-β, respectively (**Figure 2.1**). Moreover, HSPC and monocyte populations had more than 6000 differentially accessible peaks (Supplemental Figure 3A) compared to the approximately 15000 and 6000 differentially accessible peaks in MCF-7 cells in response to high dose retinoic acid and TGF-β, respectively (Figure 1A).

Next, we annotated the location of peaks based on where in the genome they were located relative to gene bodies and quantified what proportion of peaks fell into annotation categories such as promoter, intergenic, exonic, intronic, etc. ATAC-seq peaks from MCF-7 cells had a larger proportion of peaks at gene promoters (within 3 kilobases upstream or downstream of the transcription start site) whereas a greater proportion of the DNase I hypersensitive sites in the HSPC and monocyte populations

were from distal intergenic regions compared to promoters (**Figures 2.9A, B**). This finding could be the result of inherent differences in the assays or could reflect biological differences. Moreover, the MCF-7 data had a greater proportion of peaks located at gene promoters, which could in principle bias our results toward having a larger degree of concordance because accessibility changes at promoters were more strongly correlated with gene expression changes than distal accessible. Despite this bias, our data demonstrate less concordance.

Given the different assays used to determine genome-wide chromatin accessibility, we realigned the DNase-seq data to the hg38 reference and examined the peaks at a 'housekeeping gene' (GAPDH), hematopoietic differentiation-specific genes (CD34, CD14) and retinoic acid and TGF-β-related genes (DHRS3, SERPINA11) to spot-check that the accessibility data were similar. Indeed, there were similar accessibility profiles for GAPDH, and appropriate differences in accessibility given the cell type of signal for the other sites, indicating the accessibility data were comparable (**Figure 2.10**). Moreover, to look at similarities in accessibility genome-wide, we calculated the intersection of the consensus peak sets from hematopoietic differentiation and MCF-7 signal response data sets. We observed that approximately 55% of peaks from hematopoietic differentiation data (DNase-seq) overlapped with peaks from the MCF-7 signal response data set (ATAC-seq). These results show that the datasets do not have systematic qualitative differences in either expression or accessibility, enabling us to compare the degree of concordance across these two systems.

In the original analysis of hematopoietic differentiation, the authors found that regulatory complexity (defined as the number of accessible regions closest to a gene's transcriptional unit) was an important discriminating factor for whether changes in

accessibility corresponded to changes in expression, with areas of high complexity showing more correspondence than those of low complexity. Hence, we similarly grouped genes from our MCF-7 dataset into high and low complexity for our comparisons. We categorized genes with more than 7 peaks assigned to them using the 'nearest approach' as 'high complexity', while genes with 7 or fewer peaks were categorized as having 'low complexity' (**Figure 2.11**). The cutoff for loci complexity was calculated by taking a tertile based approach (González et al, 2015) and calling any number of peaks above the highest tertile cutoff as high and any peak below that as low complexity (**Figure 2.12**). Because high complexity genes on average had higher levels of expression in the hematopoietic differentiation data, we sought to determine if there was any difference in expression between high and low complexity genes in our MCF-7 data. The median expression of high complexity loci was similarly higher than low complexity loci in response to both exposure to high dose retinoic acid (23.30 versus 13.27 TPM) and high dose TGF-β (24.06 versus 13.05 TPM) (**Figure 2.13A**, p-value < 2.2x10-16 for both, Kolmogorov-Smirnov test) demonstrating that high complexity genes are more highly expressed as in the hematopoietic differentiation data. Despite this difference in expression, the distributions of peak widths for peaks of high and low complexity genes were similar (**Figure 2.13B**).

We began our analysis by focusing on the high complexity genes. To determine the concordance between gene expression changes and chromatin accessibility changes, we used the 'nearest approach' to assign peaks to genes. For each gene we compared the log2 of the fold change in expression between conditions versus the proportion of peaks that were differentially accessible in the same direction (i.e., peaks that increase in accessibility for genes that increase in expression after exposure to signal and vice versa).

We observed that for hematopoietic differentiation, the 100 most highly expressed high complexity genes in the HSPC and monocyte populations had a high proportion of peaks which were differentially accessible in the concordant direction, reproducing the conclusions of González et al. that large changes in expression were consistently associated with concordant changes in chromatin accessibility (**Figure 2.14A**). Next, we used this approach on our data to compare expression and accessibility changes between ethanol vehicle control and high dose retinoic acid or TGF-β. For both signals, we observed two distinct groups of genes within the top 100 most differentially expressed genes. One group of genes ('accessibility-concordant genes') behaved similarly to those in the hematopoietic differentiation data, demonstrating a concordance between expression and accessibility changes (**Figures 2.14B,C**). However, the other group of genes ('accessibility-non-concordant genes') had large expression changes with little to no peaks nearby changing in accessibility, creating a skew in the distribution toward a lower proportion of peaks being differentially accessible in a concordant manner compared to the hematopoietic differentiation data (**Figures 2.14A-C**, density plots).

Adjusting the minimum peak coverage parameter changes the number of differential peaks and the proportion of differential peaks that change in the corresponding direction of expression. We wondered if a lower minimum coverage threshold changed the qualitative result we noticed before and thus conducted the same analysis using a lower minimum peak coverage threshold for determining differential peaks (see methods). We observed that a similar pattern occurred in high complexity genes with this set of parameters (**Figures 2.15A, B**).

González and colleagues showed that for some low complexity genes, large changes in expression were not accompanied with concordant changes in accessibility

(González et al, 2015). We similarly wanted to confirm whether this decreased correspondence was the case in our data in response to retinoic acid and TGF-β. Using the same approach as before, we compared the log2 of the fold change in expression of low complexity genes to the proportion of peaks with differential accessibility in the concordant direction. The hematopoietic differentiation and signaling data for low complexity all qualitatively had genes whose expression increased without concordant changes in accessibility (**Figures 2.16A-C**). The distribution of the proportion peaks that were differentially accessible in the concordant direction for the top 100 up and downregulated genes was roughly uniform when comparing HSPCs to monocytes (**Figure 2.16**). By comparison,the distribution was skewed toward more genes having a lower proportion of peaks being differentially accessible in the concordant direction in response to signals in MCF-7 cells, especially in the case of TGF-β (**Figure 2.16**, density plots on right). Thus, while both the signaling in MCF-7 and hematopoietic data demonstrated large gene expression changes without concordant changes in chromatin accessibility with low complexity genes, a greater proportion of genes did so in the signaling data.

2.2.4 Peaks nearby genes with high concordance have lower accessibility prior to exposure to signal

We wondered what the differences were between genes that were differentially expressed and had large accessibility changes versus those that were differentially expressed and had low accessibility changes. First, for high dose retinoic acid and TGF-β, we split genes into four groups based on whether they were differentially expressed and the proportion of peaks assigned to them using the 'nearest' method that were differentially accessible in the appropriate direction. These four groups were (1) genes with differentially

upregulated expression and concordant accessibility changes (2) genes with differentially upregulated expression non-concordant accessibility changes (3) genes with differentially downregulated expression and a concordant accessibility changes, and (4) genes with with differentially downregulated expression and non-concordant accessibility changes (**Figures 2.17A,B**). We quantified the distribution of peak complexity across these groups and observed that they were similar across all four gene subgroups (**Figures 2.18A,B**).

We first asked whether the change in accessibility between these two gene groups was due to differences in the preexisting accessibility of peaks for these genes. Indeed, we found the baseline accessibility of peaks for genes with concordant increases in expression and accessibility in ethanol vehicle conditions was lower than those of peaks of genes that increase in expression without a commensurate change in chromatin accessibility (**Figure 2.18C**). This relationship was also recapitulated for concordant peaks that increase in expression and accessibility in response to high dose TGF-β (**Figure 2.18D**). Similarly, when comparing genes that are differentially downregulated in expression a similar pattern holds true in the opposite direction (**Figures 2.17C,D**, **Figures 2.18C,D**). One explanation may be that genes whose nearby chromatin was already accessible were permissive toward the action of the appropriate transcription factors to modulate expression. An alternative explanation is that the ATAC-seq assay itself had saturated in its ability to measure chromatin accessibility. In contrast, the difference in accessibility decreased between genes with a low proportion of peaks that were differentially accessible and genes with a high proportion of accessible peaks after exposure to signal (**Figures 2.18C,D**). Thus, the difference in the proportion of

accessible peaks nearby the two groups of genes was partially explained by the pre-existing chromatin accessibility.

2.2.5 Multiple approaches to integrating chromatin accessibility and gene expression changes show a low degree of concordance during signaling.

Finally, we measured to what degree the change in accessibility of chromatin nearby a gene is reflected in the change in gene expression. Because linear distance is not always a good predictor of what accessible regions interact with what genes, we used multiple approaches to assign peaks to genes. First, we used the 'nearest approach' to create a one-to-one mapping between accessible sites and genes by assigning them to the nearest transcriptional start site (Nair et al, 2021; Li et al, 2012), again comparing our signaling dataset to the hematopoietic differentiation dataset. Because many genes have multiple peaks assigned to them, we used two methods for collapsing peak values per gene: either the median accessibility of peaks across genes or the maximum (**Figure 2.19A**, schematic). We observed a stronger correlation between accessibility and expression changes in differentiation data (median approach Pearson's r = 0.34, maximum approach Pearson's r = 0.26) than in MCF-7 in response to signal (retinoic acid: median approach Pearson's r = 0.27, maximum approach Pearson's r = 0.10; TGF-β: median approach Pearson's r = 0.27, maximum approach Pearson's r = 0.10; **Figure 2.19A**, right side).

Next, we used a window-based approach where there was the possibility of a many-to-one mapping of peaks to genes. We assigned all peaks within a 100 kilobase window (Sanford et al, 2020) in order to maximize the number of differential peaks assigned to a gene (**Figures 2.20A,B**). Similar to the 'nearest' approach, we collapsed values using median accessibility change across all peaks assigned to a gene as well as

maximum accessibility per gene (**Figure 2.19B**, schematic) We observed a similar effect using this approach where there was a stronger correlation between change in accessibility and change in expression between HSPC versus monocyte versus MCF-7 cells exposed to signal (**Figure 2.19B**). Of note, the correlation coefficients were similar between both methods of assigning peaks.

We also wondered if the correlation between the extent of chromatin accessibility changes and gene expression changes would be different at the two lower doses. We used both the median and maximum peak value per gene while assigning peaks to genes using the nearest and window approaches. We observed similarly weak correlation as high dose signal using all methods at both low and medium doses (**Figures 2.20C,D**). Consequently, the correlation between the magnitude of change in gene expression and chromatin accessibility was modest across the range of doses of signals.

To see if peaks in specific genomic regions (promoters, parts of the gene body, downstream and intergenic areas) had unique relationships between change in chromatin accessibility and change in gene expression, we subsetted our correlation analysis. We annotated peaks using ChiPseeker (Yu et al, 2015) to categorize them as being at promoters, within the gene body (5' UTR, 3' UTR, intronic, and exonic sequences), downstream of the gene end, or at intergenic sequences. We used peaks assigned to genes using the 'nearest' approach and took the median change in accessibility per gene. The strongest correlation between changes in accessibility and gene expression across sets of comparisons was at promoter peaks (**Figure 2.19C**). While promoter correlation is quantitatively stronger, the overall qualitative conclusion remains the same. Thus, despite using a variety of approaches for both assigning peaks to genes as well as collapsing the accessibility of all peaks for a given gene to a single

value, we failed to appreciate a strong relationship between changes in accessibility and changes in gene expression.

Finally, we wondered if peaks that contained the motifs of transcription factors that are associated with retinoic acid and TGF-β signaling only (as opposed to all peaks) would show a stronger correlation between the changes in chromatin accessibility and gene expression. We annotated peaks with a log-likelihood score of a given motif being found in that peak and subsetted on those peaks with a nonzero log-likelihood score to examine the correlation between changes in accessibility and gene expression. Using this approach, we examined log-likelihood scores for motifs associated with retinoic acid signaling (RARA-α, HOXA13, and FOXA1) and motifs associated with TGF-β (SMAD3, SMAD4, and SMAD9). We observed that focusing on peaks annotated with peaks we would a priori expect to be involved in modulating gene expression in response to signal showed limited correlation between changes in chromatin accessibility and changes in gene expression (**Figure 2.21**).

2.3 Discussion

Here, we integrated tandem, genome-wide chromatin accessibility and transcriptomic data to characterize the extent of concordance between them in response to inductive signals. We demonstrated that while certain genes have a high degree of concordance of change between expression and accessibility changes, there is also a large group of differentially expressed genes whose local chromatin remains unchanged. By comparison, data from cell types along the hematopoietic differentiation trajectory had a much higher degree of concordance between genes with large gene expression changes and chromatin accessibility changes.

What might explain the lack of concordant changes in chromatin accessibility? One explanation could be that pre-existing chromatin accessibility dictates the de novo binding of transcription factors, but that the binding of transcription factors to those regions does not result in further changes to accessibility. Such effects have been reported in the context of glucocorticoid signaling, in which the glucocorticoid receptor almost exclusively binds to chromatin that is already accessible in response to dexamethasone (John et al, 2011). Indeed, we demonstrated that genes that lacked concordance between changes in chromatin accessibility and gene expression were more likely to have nearby chromatin that was already accessible (Figures 3C,D). It is possible that in MCF-7 cells, the transcriptional effects of RA and TGF-β do not lead to a significant change in the activity of pioneer transcription factors, which are able to bind directly to condensed or inaccessible chromatin to facilitate its opening (Zaret, 2020). Also, implicit in our approach is the assumption that an increase in accessibility is associated with an increase in expression, which is not necessarily the case if a genomic locus becomes accessible to a repressive factor or a bound repressive factor is displaced by a nucleosome.

We looked at MCF-7 cells exposed to retinoic acid and TGF-β because these two signals induce a robust transcriptional response through distinct mechanisms. RAR-α remains bound to DNA and interacts with transcriptional activators in response to retinoic acid binding, while SMAD family members require TGF-β to bind to surface receptors to translocate to the nucleus. Yet, despite these differences, we observed that many genes changed expression independent of changes in chromatin accessibility for both signals. It is, however, possible that signaling molecules that exert their effects through very different types of transcription factors may have a different profile of

concordance between changes in accessibility and gene expression. It is possible that other types of factors in a different context (e.g., different cell line) may yield a stronger correspondence.

Our data characterized molecular changes resulting from a single input (retinoic acid or TGF-β) in a clonal cell line, whereas the majority of work reporting a stronger concordance between simultaneous measurements of accessibility and transcription compared entirely different cell types or cells undergoing a directed differentiation protocol. What we have observed in the case of a single perturbation applied to cells that are not thought to change type per se is increased or decreased transcription with less concomitant nearby change in accessibility. How can one reconcile these observations? One possibility is that if we were to leave the signal on for longer, or combine it over time with the effects of several other signals, that we eventually would observe many further changes in accessibility proximal to a gene, concordant with the aforementioned results from comparisons between cell types. Whatever the source, these further changes in accessibility do not seem to occur randomly, given that they largely reflect the direction of change in transcription (increased accessibility for upregulation, decreased for downregulation). It may be that these subsequent changes in accessibility do not explicitly change transcription, but rather alter the underlying regulatory logic of the gene; i.e., the removal of a signal may not lead to a decrease in the gene's transcription, or the gene's transcription may be sensitized or desensitized to some other set of transcription factors.

2.4 Contributions

This chapter contains direct quotes and figures from Kiani *et al.* published in 2022 in BioRxiv [in revision, *Molecular Systems Biology*] (Kiani et al. 2022). We are greatly

indebted to Professor Christina Leslie and Alvaro González for many insightful discussions and for assistance in working with their datasets. We also thank the members of the Raj lab for valuable feedback, especially Ally Coté and Lee Richman.

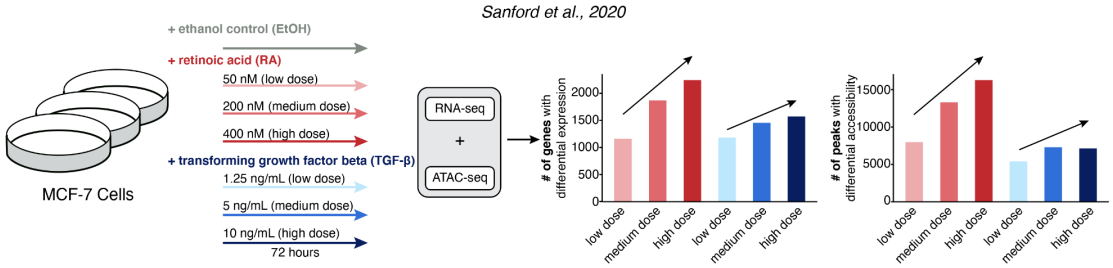

Figure 2.1 Schematic of tandem RNA-seq and ATAC-seq data.

Cells were treated with either ethanol vehicle control (gray) or three different doses of retinoic acid (shades of red) or TGF-β (shades of blue). After 72 hours of continuous exposure, bulk RNA-seq and ATAC-seq were performed on samples. We show the number of differentially expressed genes and differentially accessible peaks for each dose of each condition compared to ethanol vehicle control.

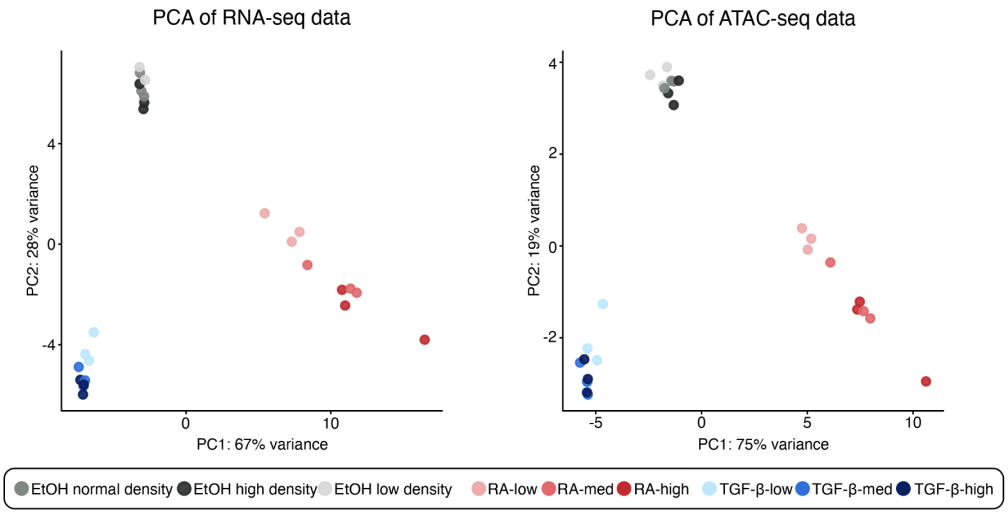

Figure 2.2 Global analysis of expression and chromatin accessibility changes in response to varying signals in MCF-7 cells.

PCA of variance stabilizing transformed raw counts from gene expression and chromatin accessibility data demonstrating the first two principal components.

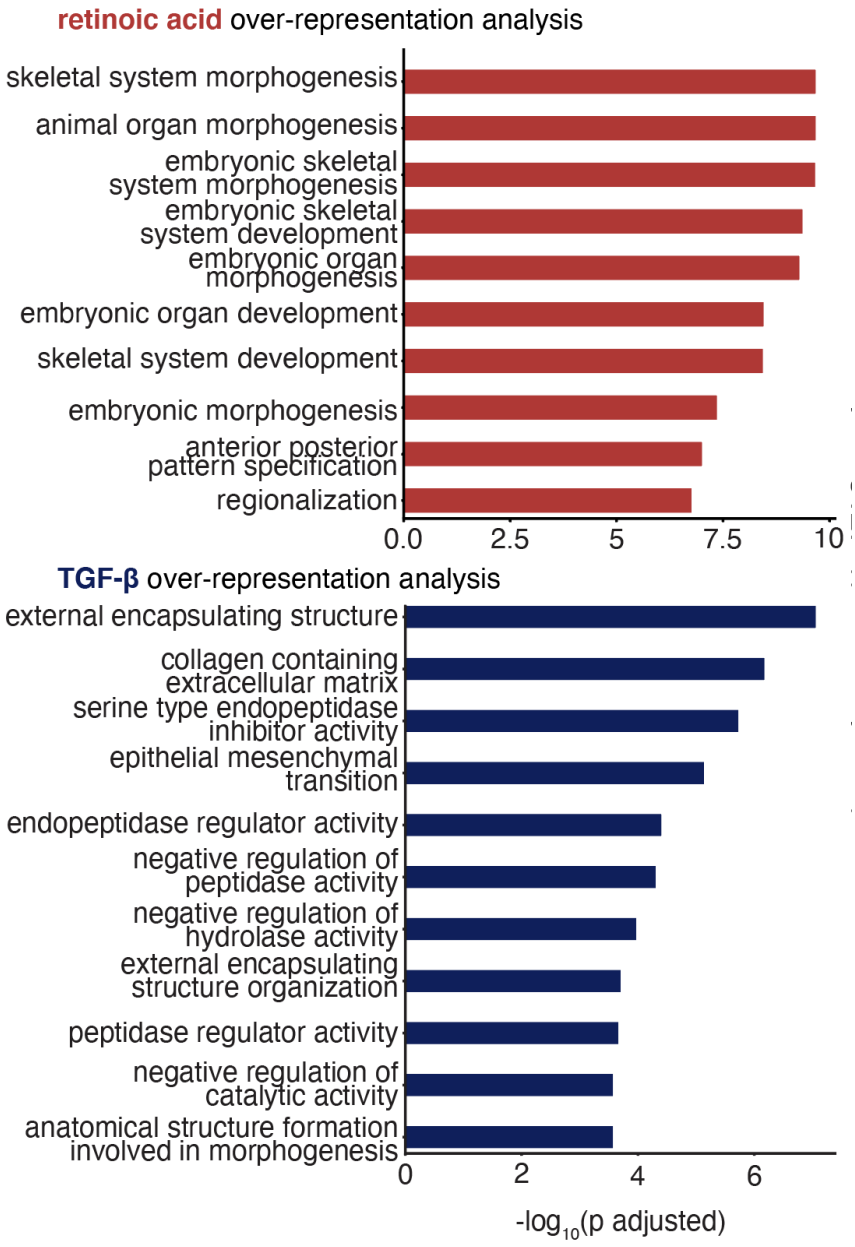
retinoic acid over-representation analysis
skeletal system morphogenesis
animal organ morphogenesis
embryonic skeletal system morphogenesis
embryonic skeletal system development
embryonic organ morphogenesis
embryonic organ development
skeletal system development
embryonic morphogenesis
anterior posterior pattern specification
regionalization
0.0
2.5
5
7.5
10
TGF-β over-representation analysis
external encapsulating structure
collagen containing extracellular matrix
serine type endopeptidase inhibitor activity
epithelial mesenchymal transition
endopeptidase regulator activity
negative regulation of peptidase activity
negative regulation of hydrolase activity
external encapsulating structure organization
peptidase regulator activity
negative regulation of catalytic activity
anatomical structure formation involved in morphogenesis
0
2
4
6
-log10(p adjusted)

Figure 2.3 Validation that changes in gene expression reflect known biology of perturbations.

Overrepresentation analysis of differentially upregulated genes in response to high dose retinoic acid (red) or TGF-β (blue). Top ten gene sets for each signal by -log10 FDR-adjusted p-value are shown.

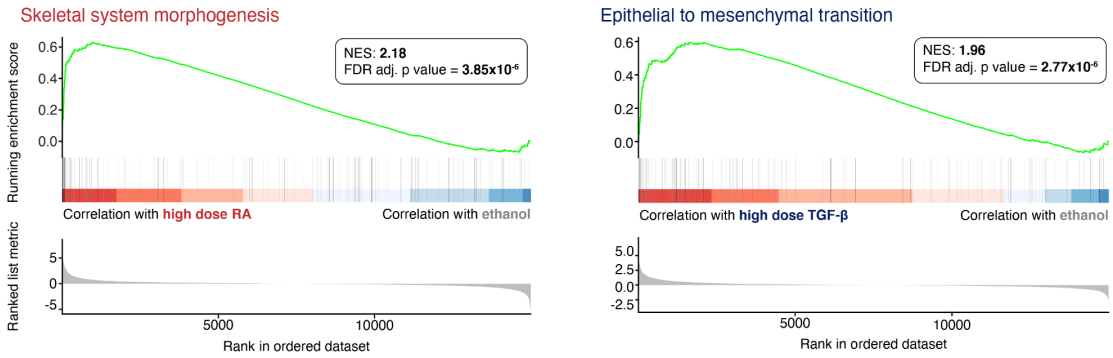

Figure 2.4 Gene set enrichment analysis of expression data further corroborates that expression changes reflect known biology of perturbations.

Gene set enrichment analysis (GSEA) (Subramanian et al, 2005) of differentially expressed genes in response to high dose retinoic acid against a gene set for skeletal system morphogenesis. Genes whose expression were differentially expressed in response to TGF-β were enriched for genes associated with epithelial-to-mesenchymal transition. Green traces represent running enrichment scores across fold change ranked gene lists.

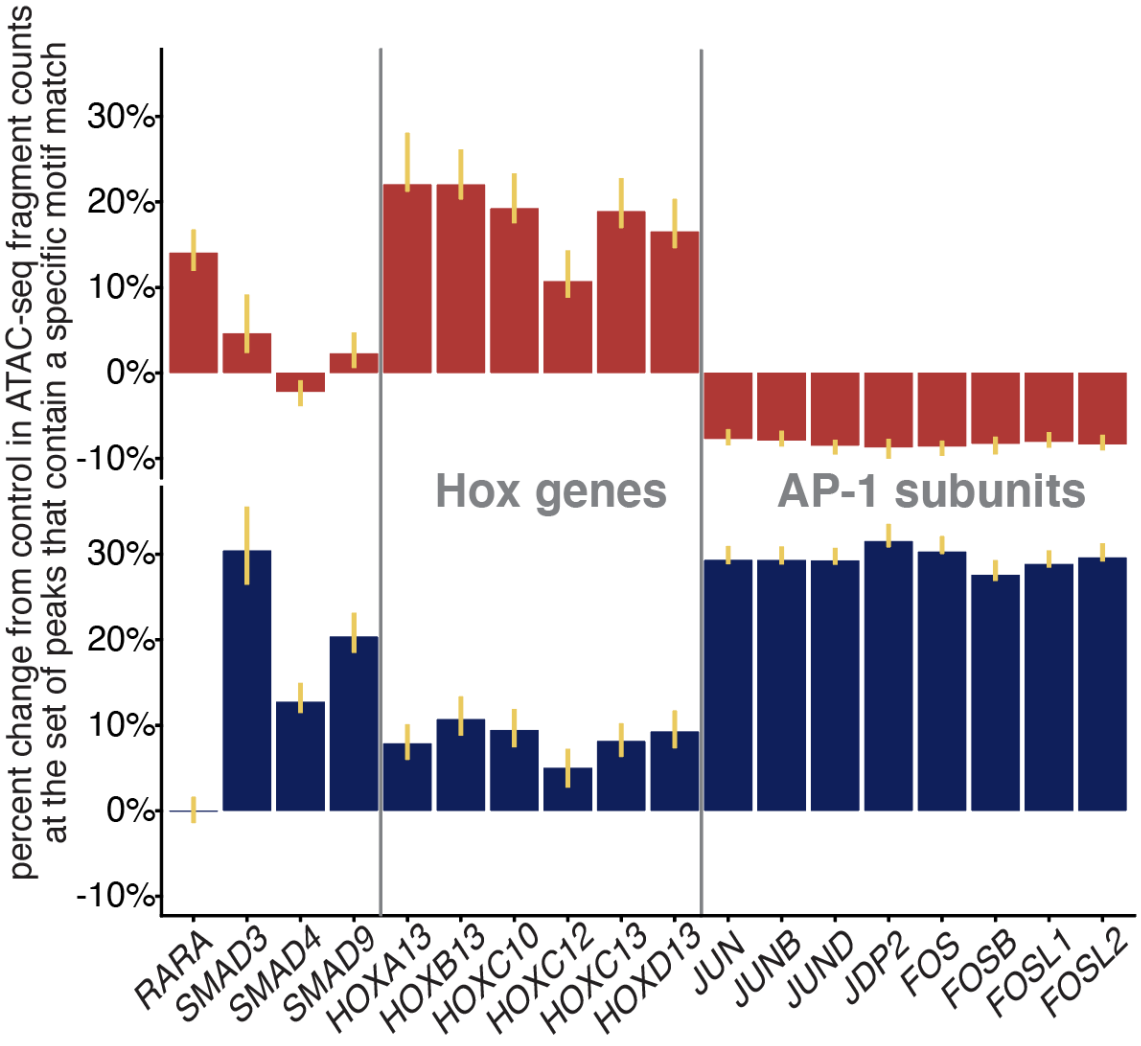

Figure 2.5 Validation that changes in chromatin accessibility reflect known biology of perturbations.

Motif enrichment analysis of differentially accessible peaks for selected motifs of transcription factors related signaling pathways of these signals. Y-axis shows percentage change of ATAC-seq signal at motif containing peaks relative to ethanol vehicle control

samples. For each condition, we pooled together replicates for all three doses. Error bars represent bootstrapped confidence intervals.

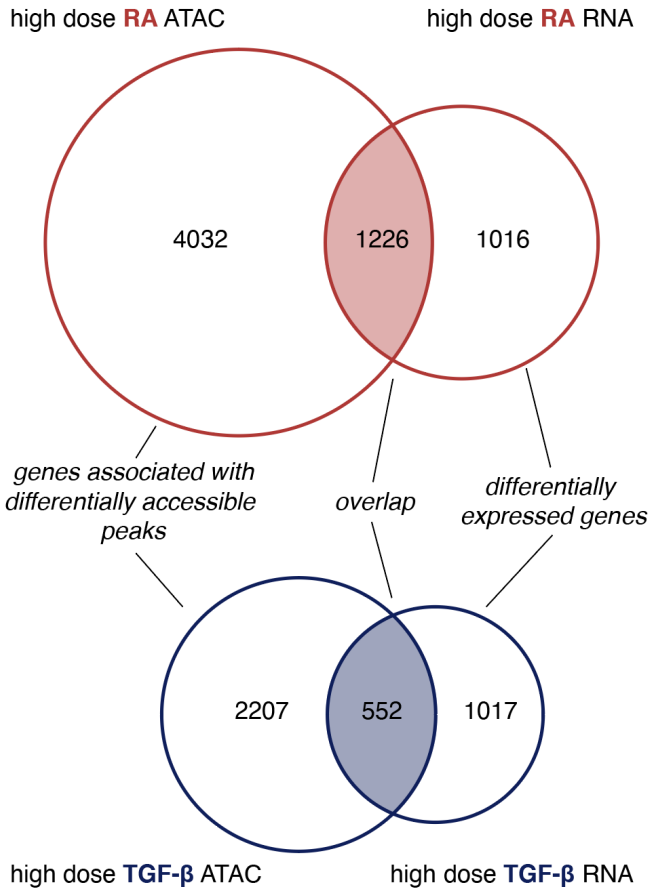

Figure 2.6 Overlap between changes in gene expression and changes in chromatin accessibility in response to high dose retinoic acid or high dose TGF-β.

Of the genes that were differentially expressed (right circle of Venn diagram) we looked at the overlap (shaded) of how many of them also had at least one differentially accessible peak (left circle). To disprove the null hypothesis that there is no association between genes that are differentially expressed and genes that have differentially accessible peaks assigned to them using the 'nearest' approach, we performed Fisher's

exact test to show the probability of these data or more extreme if the null hypothesis was true for both signals was less than 2.2x10-16.

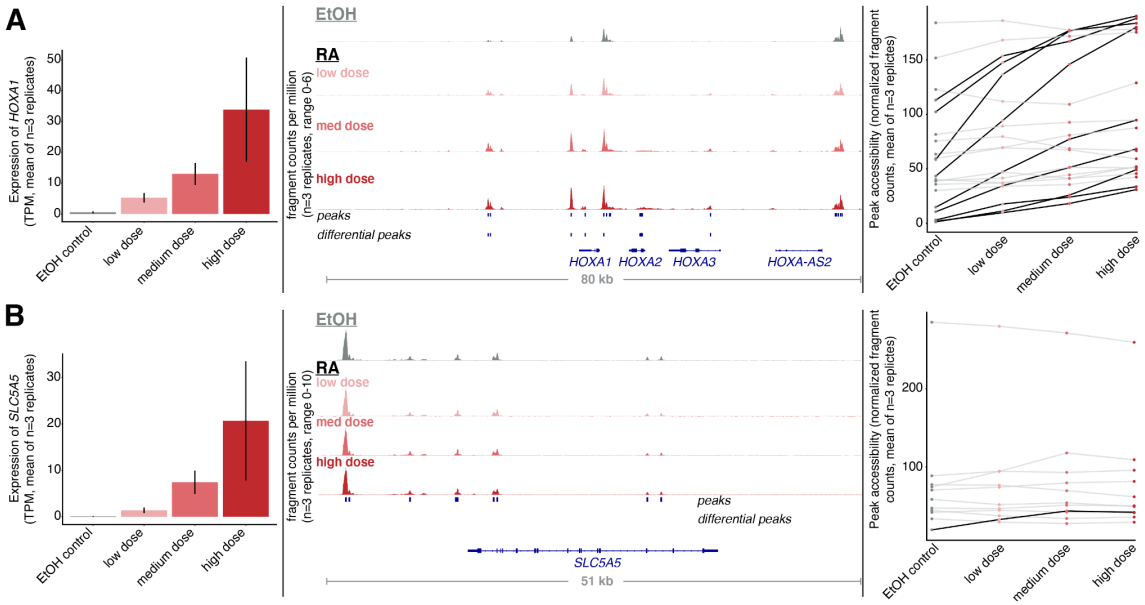

Figure 2.7 Expression and accessibility change of *HOXA1* and *SLC5A5* in response to increasing doses of retinoic acid.

(A) Left: Expression (TPM, triplicate average) in response to increasing dose of retinoic acid (error bars represent SEM). Middle: track view of HOXA1 locus with accessibility in fragments per million and peaks and differential peaks annotated. Right: quantification of peak accessibility (normalized fragment counts, triplicate average) within a 50 kilobase window of HOXA1 locus with peaks that are differentially accessible between ethanol vehicle control and high dose retinoic acid conditions marked with black lines.

(B) Left: Expression (TPM, triplicate average) in response to increasing dose of retinoic acid (error bars represent SEM). Middle: track view of SLC5A5 locus with accessibility in fragments per million and peaks and differential peaks annotated. Right: quantification of peak accessibility (normalized fragment

counts, triplicate average) within a 50 kilobase window of SLC5A5 locus with peaks that are differentially accessible between ethanol vehicle control and high dose retinoic acid conditions marked with black lines.

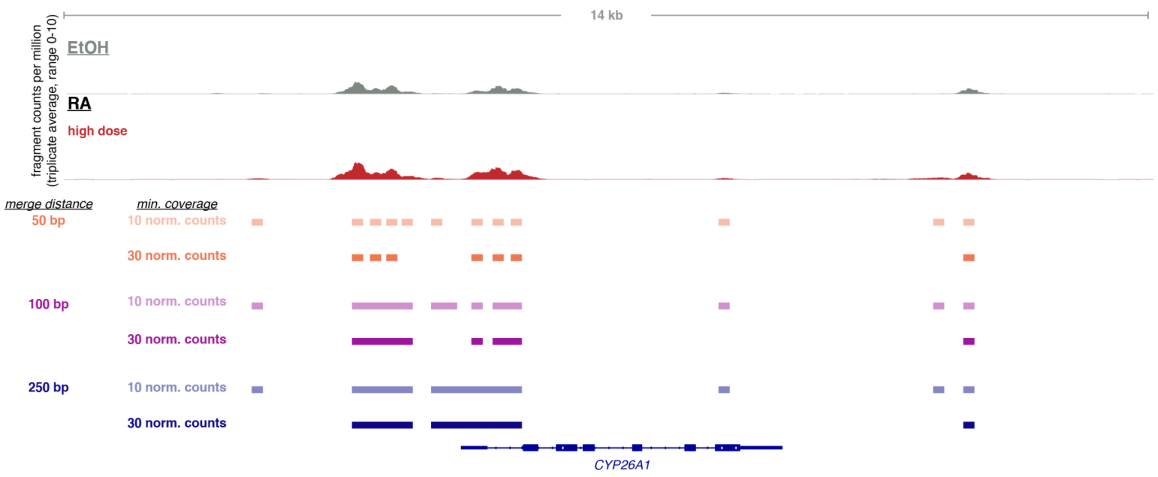

Figure 2.8 Tuning peak calling parameters

Representative peak calls at the CYP26A1 using different peak merge parameters (colors) and minimum normalized fragment count coverage (shades of the same color). Based on these results we selected a merge distance of 50 base pairs and a minimum coverage of 30 normalized fragment counts.

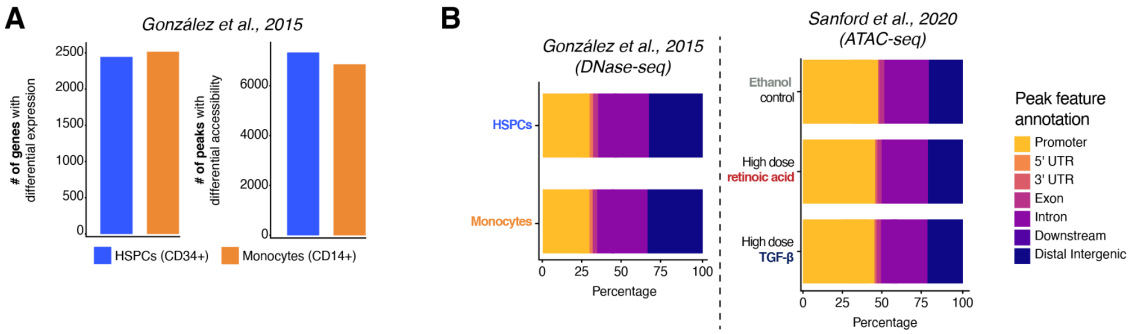

Figure 2.9 Comparison of accessibility data from hematopoietic differentiation and MCF-7 cells in response to signal.

(A) Number of differentially expressed genes (left) specific to CD34+ hematopoietic stem and progenitor cells (HSPCs, blue) and CD14+ monocytes (orange) from data from González et al., 2015 and the number of differentially accessible peaks (DNase-seq) between the two populations (right).

(B) Annotation of distribution of peak location in relation to gene transcriptional units for consensus files for HSPCs and monocytes (left). Distribution of accessible peak features for consensus peaks for MCF-7 cells in ethanol, high dose retinoic acid, and high dose TGF-β.

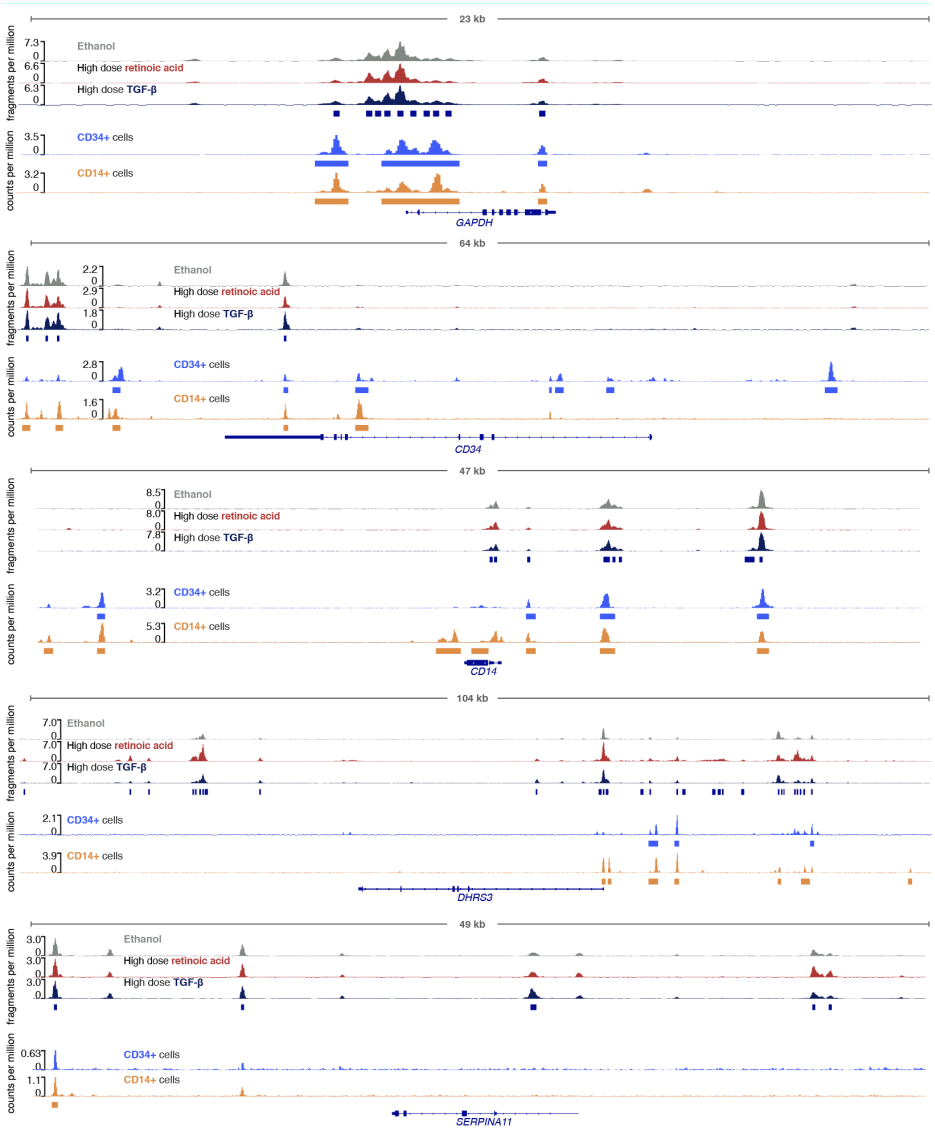

Figure 2.10 Comparison of peak calls at multiple loci across both hematopoietic differentiation and MCF-7 genome-wide accessibility data sets.

Consensus peak calls for MCF-7 signal samples (ATAC-seq) and hematopoietic differentiation samples (DNase-seq) at a 'housekeeping' gene GAPDH, hematopoietic cell-specific marker loci CD34 and CD14, a retinoic acid responsive site, DHRS3, and a

TGF-β responsive site SERPINA11. Values are fragments per million for ATAC-seq samples and counts per million for DNase-seq samples.

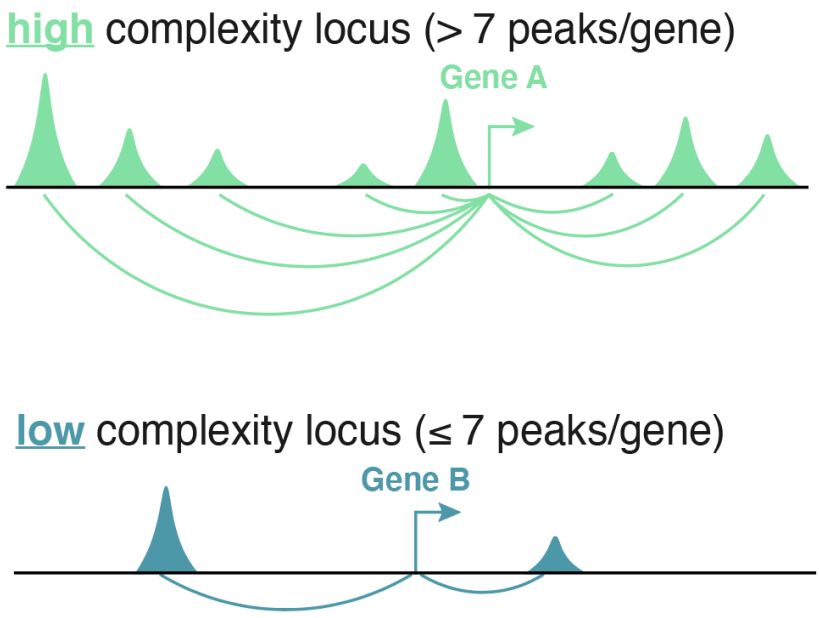

Figure 2.11 Schematic demonstrating classification of genes into "high" versus "low" complexity genes based on the number peaks assigned to a gene using the 'nearest' approach.

High complexity genes (light green) are characterized by greater than 7 peaks assigned to a given gene by the 'nearest' approach while low complexity genes (teal) have 7 or fewer genes assigned to them.

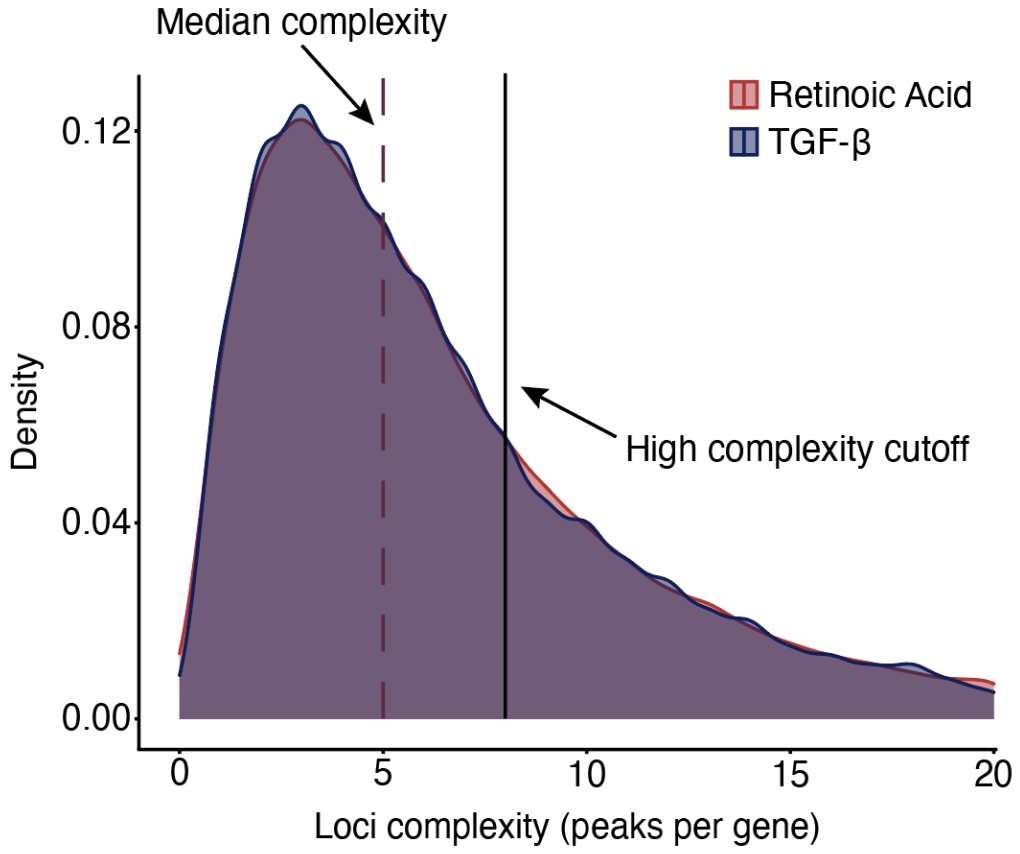

Figure 2.12 Distribution of gene complexity in response to high doses of both perturbations.

Density plot of number of peaks per gene in retinoic acid (red) and TGF-β (blue, overlap in purple) with median complexity marked by dotted line and high complexity cutoff marked by solid line.

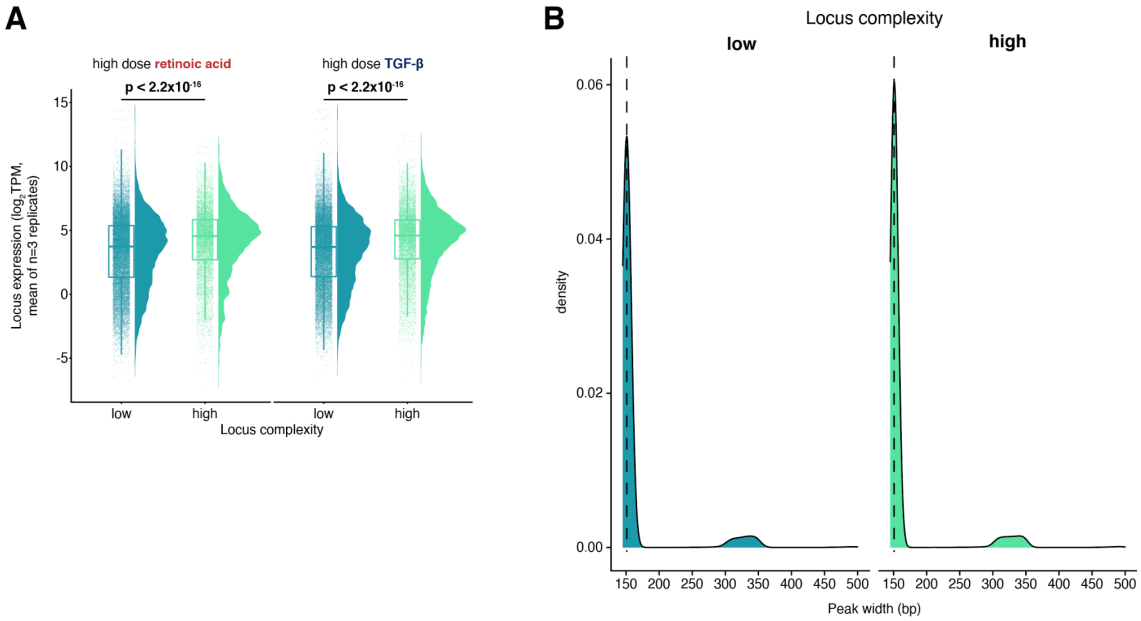

Figure 2.13 Expression and peak width distributions in MCF-7 signal data based on locus complexity.

(A) log2-transformed expression of low complexity (teal) and high complexity genes (green) in response to retinoic acid (left) and TGF-β (right). P-values represent the probability of these data or more extreme under the null hypothesis that the distribution of gene expression values were drawn from the same probability distribution via the Kolmogorov-Smirnov test.

(B) Distribution of peak widths for low complexity (teal) and high complexity (green) peaks with the median peak width (151 base pairs) marked by the dotted black line.

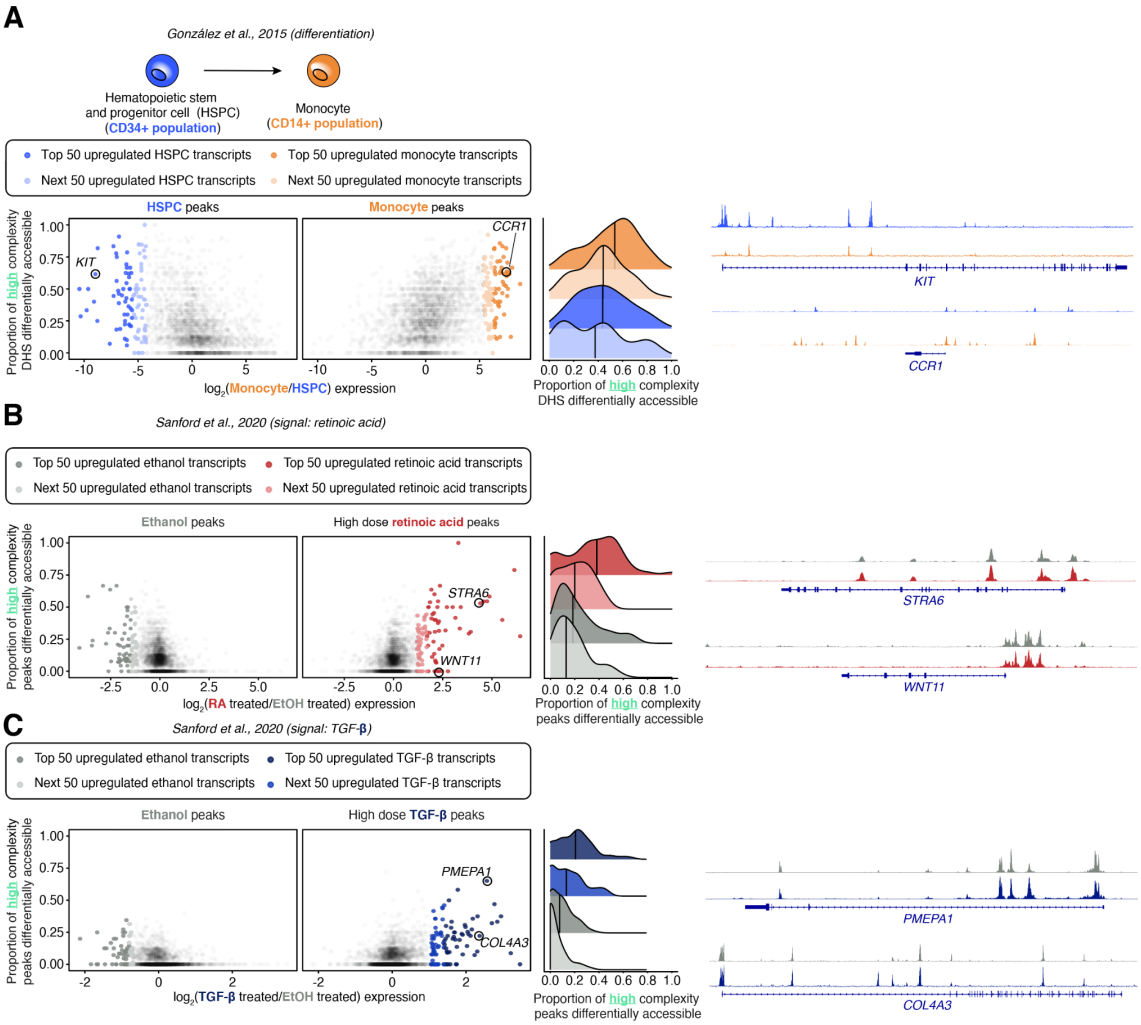

Figure 2.14 Signaling shows less concordance between highly differentially expressed genes and chromatin accessibility changes compared to hematopoietic differentiation data for high complexity genes.

(A) Concordance between expression and accessibility changes between hematopoietic stem and progenitor cells and monocytes. Left: plot showing changes in gene expression in CD34+ hematopoietic stem and progenitor cells (blue) and CD14+

monocytes (orange) from González et al., 2015 (schematic, top). For the plots, each dot is a gene, and on the x axis is log2 fold change in expression and on the y-axis the proportion of differentially accessible DHSs for each associated gene. The top 100 most highly expressed genes in hematopoietic stem and progenitor cells and monocytes are colored in shades of orange and blue, respectively. Middle: density plot of the distribution of the proportion of high complexity DHS associated with the top 100 expressed genes in CD34+ hematopoietic stem and progenitor cells and CD14+ monocytes with median value marked by vertical black line. Right: example tracks DNase I sequencing data for KIT and CCR1 (marked on plot on left).

(B) Concordance between expression and accessibility changes between cells exposed to ethanol vehicle control and high dose retinoic acid. Left: plot showing changes in gene expression and chromatin accessibility between ethanol vehicle control and high dose retinoic acid. Each dot is a gene, and on the x axis is the log2 fold change in expression and on the y-axis the proportion of differentially accessible ATAC-seq peaks for each gene. The top 100 most highly expressed genes in ethanol vehicle control and high dose retinoic acid are colored in shades of gray and red, respectively. Middle: density plot of the distribution of the proportion of high complexity ATAC-seq peaks associated with the top 100 expressed genes in ethanol vehicle control and high dose retinoic acid with median value marked by vertical black line. Right: example ATAC-seq tracks of STRA6 and WNT11.

(C) Concordance between expression and accessibility changes between cells exposed to ethanol vehicle control and high dose TGF-β. Left: plot showing changes in gene expression and chromatin accessibility between ethanol vehicle control and high dose TGF-β. Each dot is a gene, and on the x axis is the log2 fold change in

expression and on the y-axis the proportion of differentially accessible ATAC-seq peaks for each gene. The top 100 most highly expressed genes in ethanol vehicle control and high dose TGF-β are colored in shades of gray and blue, respectively. Middle: density plot of the distribution of the proportion of high complexity ATAC-seq peaks associated with the top 100 expressed genes in ethanol vehicle control and high dose retinoic acid with median value marked by vertical black line. Right: example ATAC-seq tracks of PMEPA1 and COL4A3.

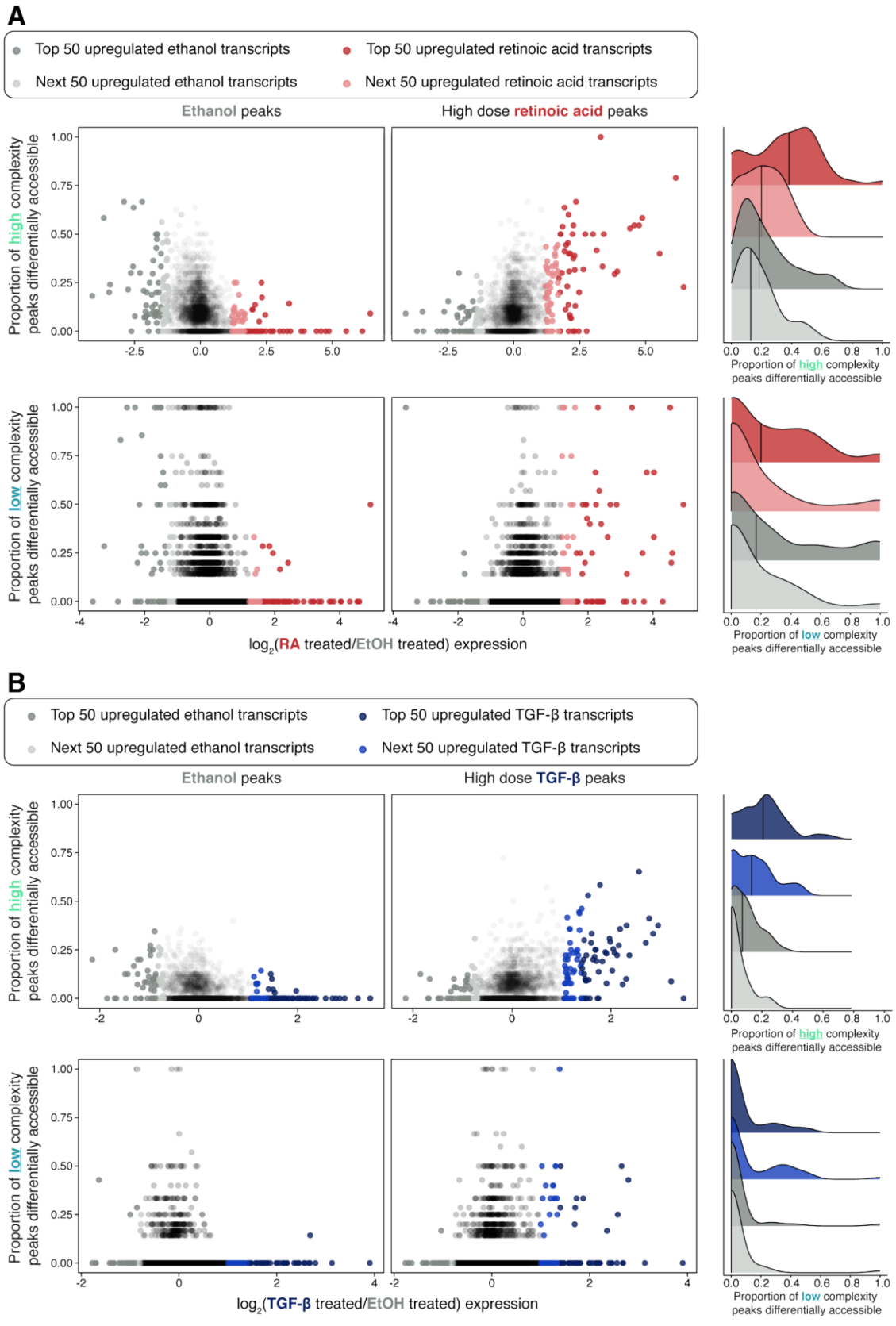
A
Top 50 upregulated ethanol transcripts
Next 50 upregulated ethanol transcripts
Top 50 upregulated retinoic acid transcripts
Next 50 upregulated retinoic acid transcripts
Ethanol peaks
High dose retinoic acid peaks
Proportion of high complexity peaks differentially accessible
Proportion of low complexity peaks differentially accessible
log2(RA treated/EtOH treated) expression
B
Top 50 upregulated ethanol transcripts
Next 50 upregulated ethanol transcripts
Top 50 upregulated TGF-β transcripts
Next 50 upregulated TGF-β transcripts
Ethanol peaks
High dose TGF-β peaks
log2(TGF-β treated/EtOH treated) expression

Figure 2.15 Concordance between gene expression change and proportion of differentially accessible peaks per gene for high and low complexity genes using a lower minimum coverage threshold for differential peaks.

(A) Concordance between expression and accessibility changes between cells exposed to ethanol vehicle control and high dose retinoic acid. Left: plot showing changes in gene expression and chromatin accessibility between ethanol vehicle control and high dose retinoic acid for high and low complexity genes. Each dot is a gene, and on the x axis is the log2 fold change in expression and on the y-axis the proportion of differentially accessible ATAC-seq peaks for each gene. The top 100 most highly expressed genes in ethanol vehicle control and high dose retinoic acid are colored in shades of gray and red, respectively. Right: density plot of the distribution of the proportion of high complexity ATAC-seq peaks associated with the top 100 expressed genes in ethanol vehicle control and high dose retinoic acid with median value marked by vertical black line.

(B) Concordance between expression and accessibility changes between cells exposed to ethanol vehicle control and high dose TGF-β. Left: plot showing changes in gene expression and chromatin accessibility between ethanol vehicle control and high dose retinoic acid for high and low complexity genes. Each dot is a gene, and on the x axis is the log2 fold change in expression and on the y-axis the proportion of differentially accessible ATAC-seq peaks for each gene. The top 100 most highly expressed genes in ethanol vehicle control and high dose retinoic acid are colored in shades of gray and blue, respectively. Right: density plot of the distribution of the proportion of high complexity ATAC-seq peaks associated with

the top 100 expressed genes in ethanol vehicle control and high dose retinoic acid with median value marked by vertical black line.

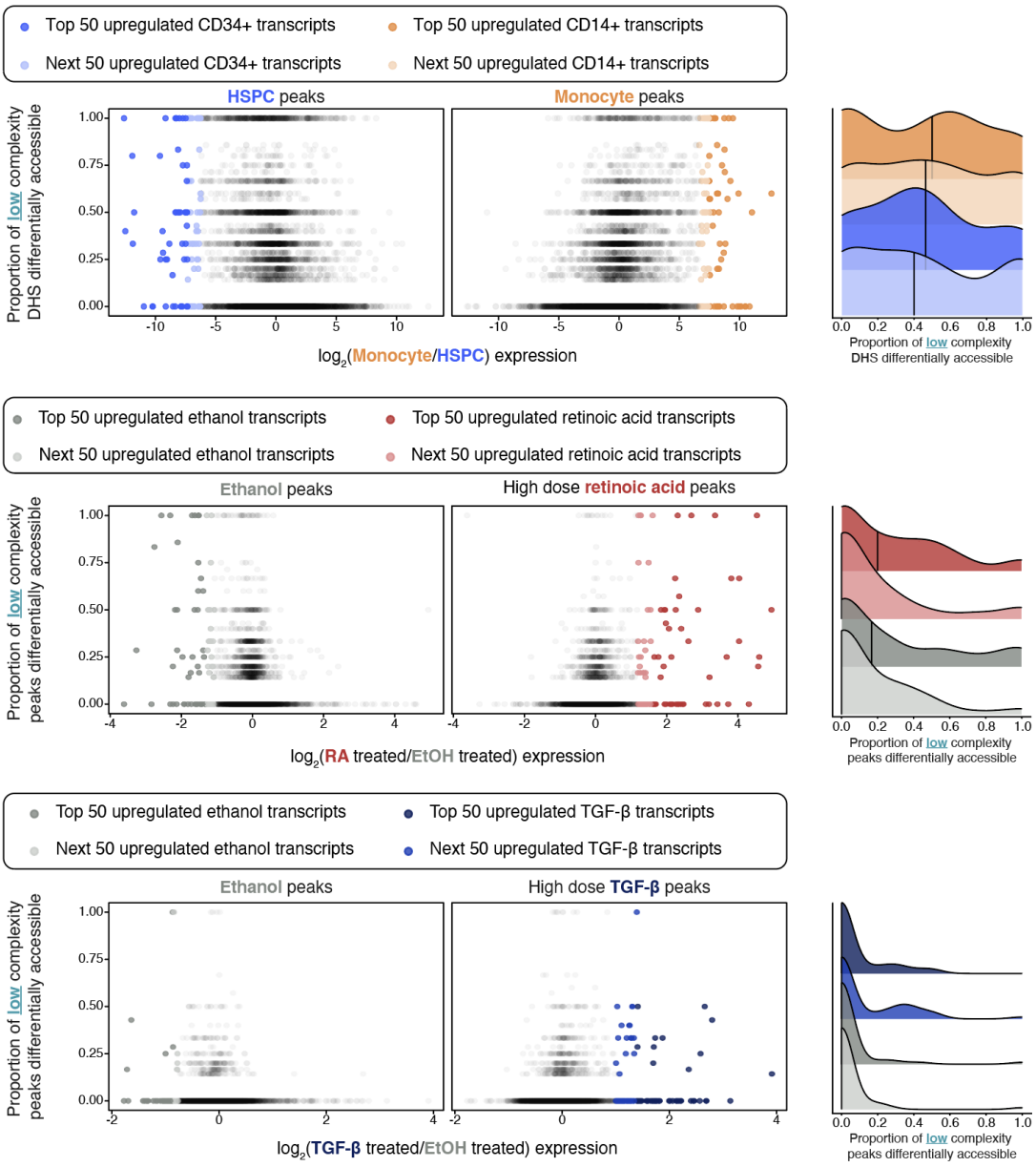

Figure 2.16 Concordance between gene expression change and proportion of differentially accessible peaks per gene for low complexity genes.

Concordance between expression and accessibility changes between hematopoietic stem and progenitor cells and monocytes, ethanol control and high dose retinoic acid, and

ethanol control and high dose TGF-β. Left: For the plots, each dot is a gene, and on the x axis is log2 fold change in expression and on the y-axis the proportion of differentially accessible DHSs/peaks for each associated gene. Right: density plot of the distribution of the proportion of high complexity DHS or peaks associated with the top 100 expressed genes in either condition with median value marked by vertical black line.

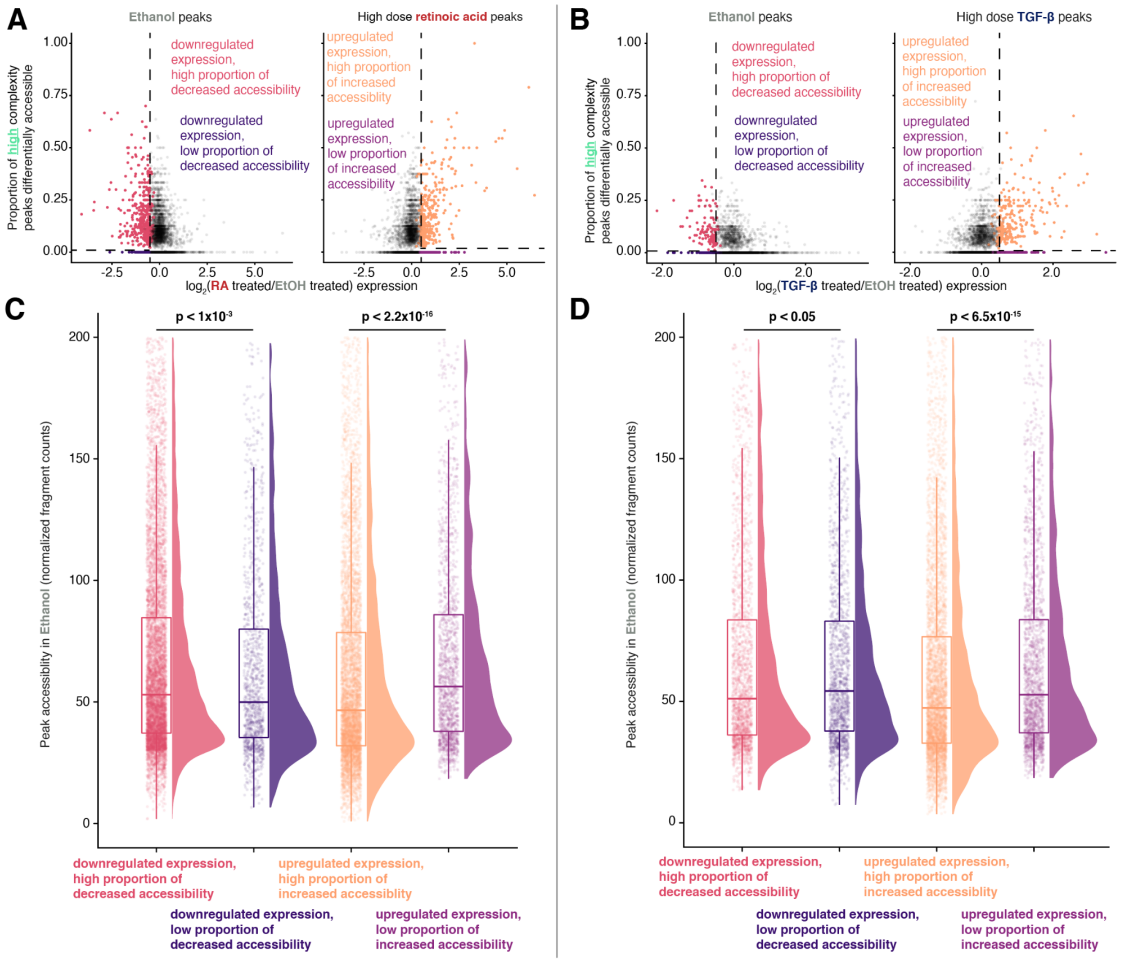

Figure 2.17 Separation of differentially expressed genes in response to signal into high and low concordance groups shows differences in pre-existing accessibility.

(A) Categorization of differentially expressed genes in response to high dose retinoic acid based on direction of expression change and proportion of peaks differentially accessible in the same direction.

(B) Categorization of differentially expressed genes in response to high dose TGF-β based on direction of expression change and proportion of peaks differentially accessible in the same direction.

(C) Differential accessibility in ethanol vehicle control conditions prior to addition of high dose retinoic acid. Accessibility of every peak assigned using the 'nearest' approach for gene groups from (a) in ethanol vehicle control conditions. P-values represent the probability of these data or more extreme under the null hypothesis that the distribution of peak accessibilities were drawn from the same probability distribution via the Kolmogorov-Smirnov test.

(D) Differential accessibility in ethanol vehicle control conditions prior to addition of high dose TGF-β. Accessibility of every peak assigned using the 'nearest' approach for gene groups from (b) in ethanol vehicle control conditions. P-values represent the probability of these data or more extreme under the null hypothesis that the distribution of peak accessibilities were drawn from the same probability distribution via the Kolmogorov-Smirnov test.

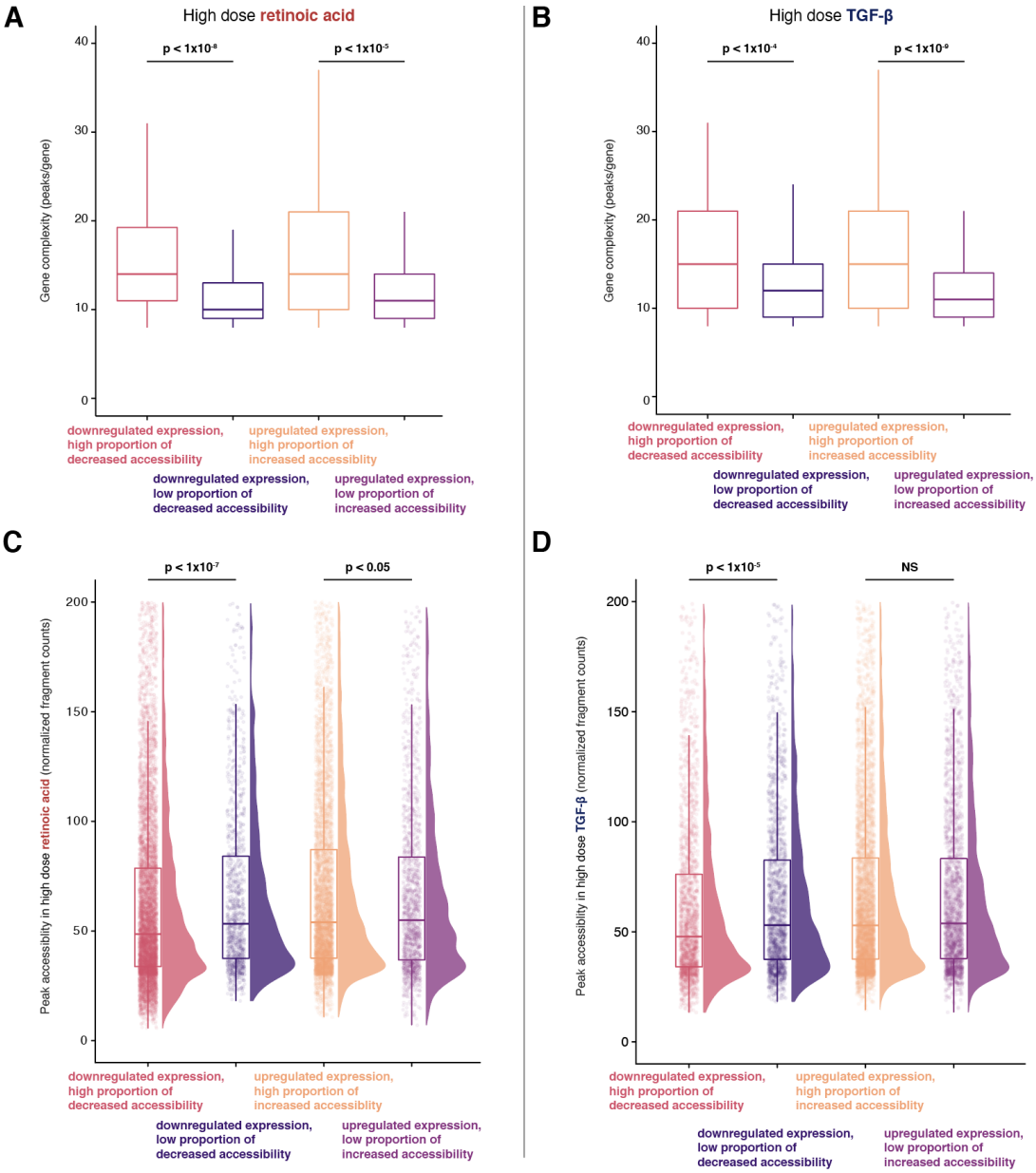
A
High dose retinoic acid
p < 1x10-8
p < 1x10-5
Gene complexity (peaks/gene)
0
10
20
30
40
downregulated expression, high proportion of decreased accessibility
upregulated expression, high proportion of increased accessiblity
downregulated expression, low proportion of decreased accessibility
upregulated expression, low proportion of increased accessibility
B
High dose TGF-β
p < 1x10-4
p < 1x10-9
Gene complexity (peaks/gene)
0
10
20
30
40
downregulated expression, high proportion of decreased accessibility
upregulated expression, high proportion of increased accessiblity
downregulated expression, low proportion of decreased accessibility
upregulated expression, low proportion of increased accessibility
C
p < 1x10-7
p < 0.05
Peak accessibility in high dose retinoic acid (normalized fragment counts)
0
50
100
150
200
downregulated expression, high proportion of decreased accessibility
upregulated expression, high proportion of increased accessiblity
downregulated expression, low proportion of decreased accessibility
upregulated expression, low proportion of increased accessibility
D
p < 1x10-5
NS
Peak accessibility in high dose TGF-β (normalized fragment counts)
0
50
100
150
200
downregulated expression, high proportion of decreased accessibility
upregulated expression, high proportion of increased accessiblity
downregulated expression, low proportion of decreased accessibility
upregulated expression, low proportion of increased accessibility

Figure 2.18 Accessibility-concordant and accessibility-non-concordant genes have similar loci complexity and differences in peak accessibility after exposure to signal depending on change in gene expression.

(A) Distribution of loci complexity the four groups of genes with differential expression in response to high dose retinoic acid.

(B) Distribution of loci complexity the four groups of genes with differential expression in response to high dose TGF-β.

(C) Accessibility after exposure to high dose retinoic acid. Accessibility of every peak assigned using the 'nearest' approach for gene groups based on accessibility concordance. P-values represent the probability of these data or more extreme under the null hypothesis that the distribution of peak accessibilities were drawn from the same probability distribution via the Kolmogorov-Smirnov test.

(D) Accessibility after exposure to high dose TGF-β. Accessibility of every peak assigned using the 'nearest' approach for gene groups based on accessibility concordance. P-values represent the probability of these data or more extreme under the null hypothesis that the distribution of peak accessibilities were drawn from the same probability distribution via the Kolmogorov-Smirnov test.

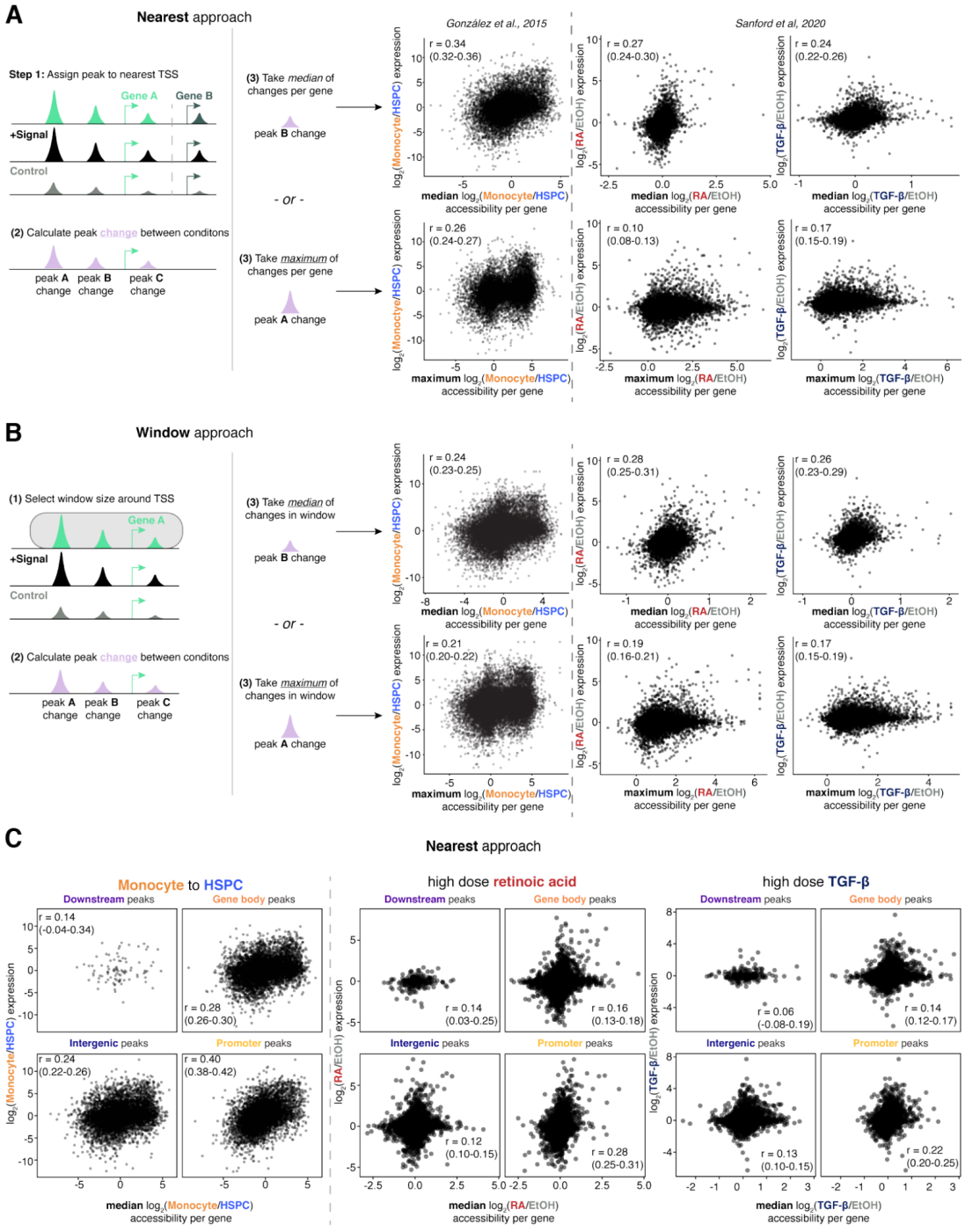
A
Nearest approach
Step 1: Assign peak to nearest TSS
Gene A
Gene B
+Signal
Control
(2) Calculate peak change between conditons
peak A change
peak B change
peak C change
(3) Take median of changes per gene
peak B change
- or -
(3) Take maximum of changes per gene
peak A change
González et al., 2015
Sanford et al, 2020
r = 0.34 (0.32-0.36)
r = 0.27 (0.24-0.30)
r = 0.24 (0.22-0.26)
r = 0.26 (0.24-0.27)
r = 0.10 (0.08-0.13)
r = 0.17 (0.15-0.19)
log2(Monocyte/HSPC) expression
log2(RA/EtOH) expression
log2(TGF-β/EtOH) expression
median log2(Monocyte/HSPC) accessibility per gene
median log2(RA/EtOH) accessibility per gene
median log2(TGF-β/EtOH) accessibility per gene
maximum log2(Monocyte/HSPC) accessibility per gene
maximum log2(RA/EtOH) accessibility per gene
maximum log2(TGF-β/EtOH) accessibility per gene
B
Window approach
(1) Select window size around TSS
Gene A
+Signal
Control
(2) Calculate peak change between conditons
peak A change
peak B change
peak C change
(3) Take median of changes in window
peak B change
- or -
(3) Take maximum of changes in window
peak A change
r = 0.24 (0.23-0.25)
r = 0.28 (0.25-0.31)
r = 0.26 (0.23-0.29)
r = 0.21 (0.20-0.22)
r = 0.19 (0.16-0.21)
r = 0.17 (0.15-0.19)
C
Nearest approach
Monocyte to HSPC
high dose retinoic acid
high dose TGF-β
Downstream peaks
Gene body peaks
Intergenic peaks
Promoter peaks
r = 0.14 (-0.04-0.34)
r = 0.28 (0.26-0.30)
r = 0.24 (0.22-0.26)
r = 0.40 (0.38-0.42)
r = 0.14 (0.03-0.25)
r = 0.16 (0.13-0.18)
r = 0.12 (0.10-0.15)
r = 0.28 (0.25-0.31)
r = 0.06 (-0.08-0.19)
r = 0.14 (0.12-0.17)
r = 0.13 (0.10-0.15)
r = 0.22 (0.20-0.25)
log2(Monocyte/HSPC) expression
log2(RA/EtOH) expression
log2(TGF-β/EtOH) expression
median log2(Monocyte/HSPC) accessibility per gene
median log2(RA/EtOH) accessibility per gene
median log2(TGF-β/EtOH) accessibility per gene

Figure 2.19 Multiple approaches to quantifying peak accessibility shows low correlation between gene expression changes and accessibility changes in signaling.

(A) 'Nearest' approach to assigning peaks to genes shows less concordance in signaling compared to hematopoietic differentiation. Left: schematic showing 'nearest' approach where peaks are assigned to the nearest transcriptional site and change in accessibility (purple) on a per-gene basis is calculated by either median change in accessibility (top row) or maximum peak change (bottom row). Right: scatter plots showing change in peak accessibility (median or maximum) versus log2 fold change in expression on y axis for hematopoietic differentiation data from González et al. (left column) and for high dose retinoic acid and high dose TGF-β (right two columns). Pearson's correlation coefficients reported with 95% confidence interval from bootstrapping with 10,000 replicates in parentheses.

(B) 'Window' approach to assigning peaks to genes shows less concordance in signaling compared to hematopoietic differentiation. Left: schematic showing 'window' approach where all peaks within a certain window of the the transcriptional start site are assigned to that gene and the change in accessibility (purple) on a per-gene basis is calculated by the median change in accessibility (top row) or the maximum change in accessibility (bottom row). Right: scatter plots showing change in peak accessibility (median or maximum) using 'window' approach with a 100 kilobase window versus log2 fold change in expression on y axis for hematopoietic differentiation data from González et al. (left column) and for high dose retinoic acid

and high dose TGF-β (right two columns). Pearson's correlation coefficients reported with 95% confidence interval from bootstrapping with 10,000 replicates in parentheses.

(C) Using 'nearest' approach to look for correlation between accessibility and gene expression changes based on annotations of peak location. First two columns showing correlation for hematopoietic differentiation data from González et al, and right four columns showing correlation for high dose retinoic acid and high dose TGF-β, respectively. Pearson's correlation coefficients reported with 95% confidence interval from bootstrapping with 10,000 replicates in parentheses.

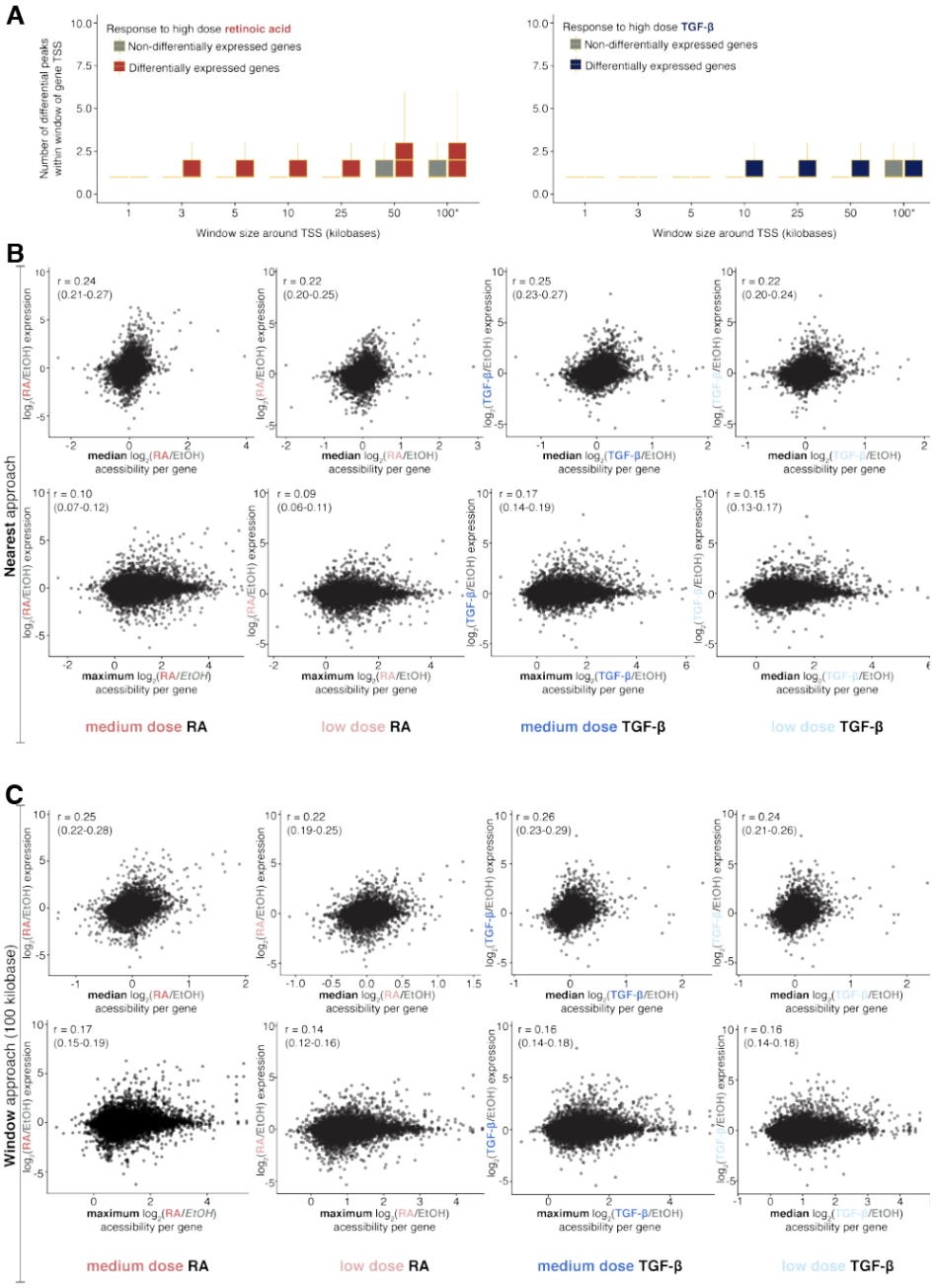
A
Response to high dose retinoic acid
Non-differentially expressed genes
Differentially expressed genes
Number of differential peaks within window of gene TSS
Window size around TSS (kilobases)
Response to high dose TGF-β
Non-differentially expressed genes
Differentially expressed genes
Window size around TSS (kilobases)
B
Nearest approach
r = 0.24 (0.21-0.27)
r = 0.22 (0.20-0.25)
r = 0.25 (0.23-0.27)
r = 0.22 (0.20-0.24)
r = 0.10 (0.07-0.12)
r = 0.09 (0.06-0.11)
r = 0.17 (0.14-0.19)
r = 0.15 (0.13-0.17)
medium dose RA
low dose RA
medium dose TGF-β
low dose TGF-β
C
Window approach (100 kilobase)
r = 0.25 (0.22-0.28)
r = 0.22 (0.19-0.25)
r = 0.26 (0.23-0.29)
r = 0.24 (0.21-0.26)
r = 0.17 (0.15-0.19)
r = 0.14 (0.12-0.16)
r = 0.16 (0.14-0.18)
r = 0.16 (0.14-0.18)
medium dose RA
low dose RA
medium dose TGF-β
low dose TGF-β

Figure 2.20 Effect of window size on number of differentially accessible peaks based on gene expression change and correlation of gene expression and accessibility changes using medium and low dose signals.

(A) Distributions of number of differentially accessible peaks for differentially expressed and non-differentially expressed genes in response to high dose retinoic acid (left) or high dose TGF-β (right) based on window size around transcriptional start site (TSS).

(B) 'Nearest' approach to assigning peaks to genes shows less concordance in signaling compared to hematopoietic differentiation. Scatter plots showing change in peak accessibility (median or maximum) versus log2 fold change in expression on y axis for medium and low dose retinoic acid (first two columns) and medium and low dose TGF-β (second two columns). Pearson's correlation coefficients reported with 95% confidence interval from bootstrapping with 10,000 replicates in parentheses.

(C) 'Window' approach to assigning peaks to genes shows less concordance in signaling compared to hematopoietic differentiation. Scatter plots showing change in peak accessibility (median or maximum) versus log2 fold change in expression on y axis for medium and low dose retinoic acid (first two columns) and medium and low dose TGF-β (second two columns). Pearson's correlation coefficients reported with 95% confidence interval from bootstrapping with 10,000 replicates in parentheses.

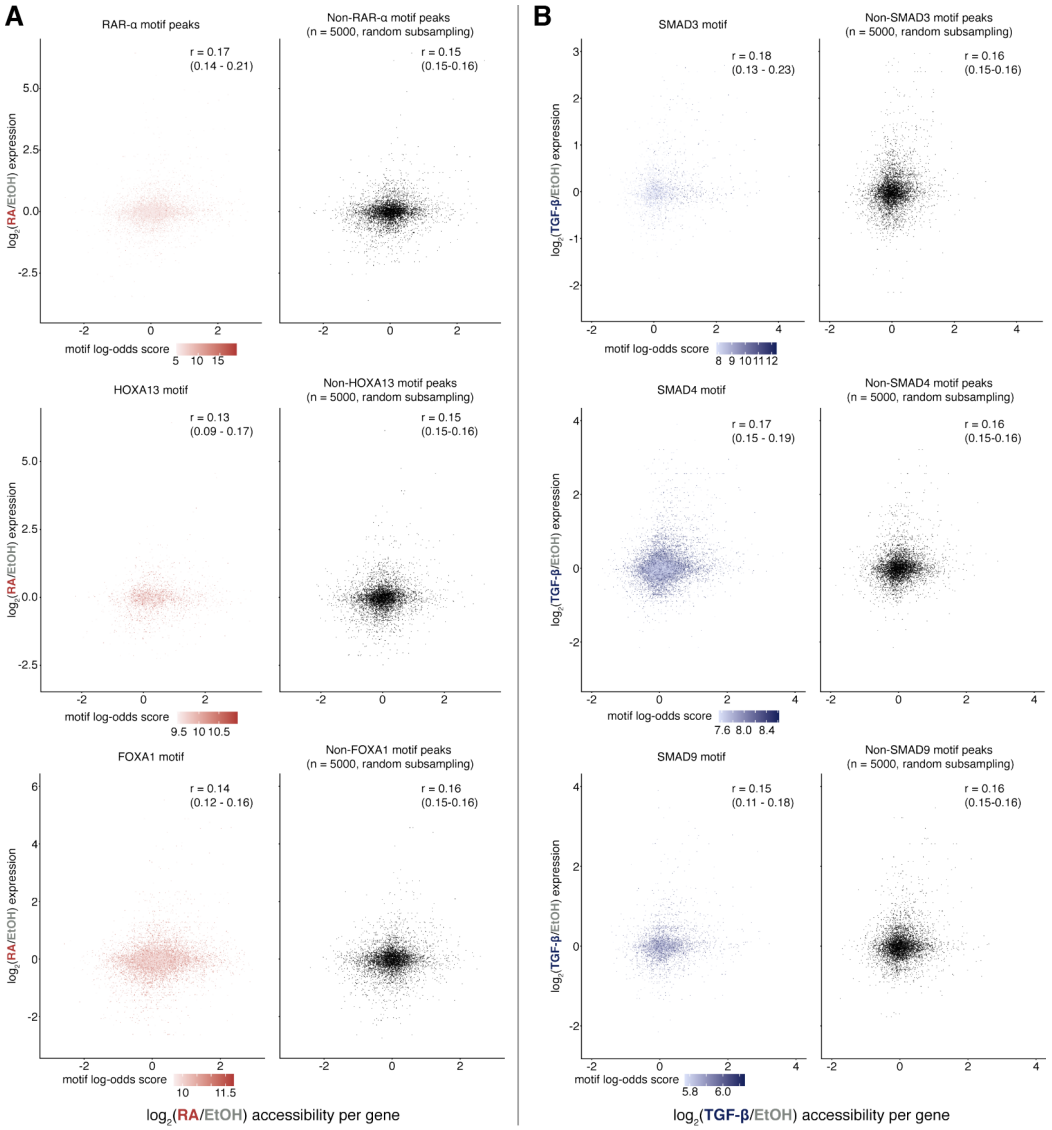

Figure 2.21 Focusing on peaks annotated for biologically relevant transcription factor motifs fails to demonstrate a strong correlation between the magnitude of gene expression and chromatin accessibility changes.

(A) Peaks annotated for motifs of transcription factors related to retinoic acid biology (RAR-α, HOXA13, FOXA1, left column) showed weak correlation between

changes in gene expression and chromatin accessibility in response to high dose retinoic acid. Peaks are colored based on the log-odds of a motif being present in a given peak. Plot of expression and accessibility change for 5000 randomly sampled peaks lacking the corresponding peak (right column). Pearson's correlation for peaks not having a given motif are for all peaks without that motif, not the 5000 subsampled peaks. Pearson's correlation coefficients reported with 95% confidence interval from bootstrapping with 10,000 replicates in parentheses.

(B) Peaks annotated for motifs of transcription factors related to retinoic acid biology (SMAD3, SMAD4, SMAD9, left column) showed weak correlation between changes in gene expression and chromatin accessibility in response to high dose TGF-β. Peaks are colored based on the log-odds of a motif being present in a given peak. Plot of expression and accessibility change for 5000 randomly sampled peaks lacking the corresponding peak (right column). Pearson's correlation for peaks not having a given motif are for all peaks without that motif, not the 5000 subsampled peaks. Pearson's correlation coefficients reported with 95% confidence interval from bootstrapping with 10,000 replicates in parentheses.

CHAPTER 3: DETERMINING DRIVERS OF A RARE, EARLY-INVADING SUBPOPULATION OF CLONAL MELANOMA CELLS

3.1 Introduction

A devastating feature of many cancers, including cutaneous melanoma, is the ability of cells to metastasize to distant sites, gain a foothold and begin rapidly dividing, and, eventually, disrupt end organ function at the site of metastasis, making metastasis a large factor in cancer morbidity and mortality. Metastasis involves rare cells in the primary tumor to undergo multiple molecular and behavioral changes to first become invasive to leave the tumor and travel via the bloodstream or lymphatics, establish itself at a new site, and then revert to a proliferative state to create the metastasis (Mittal 2018; Mani et al. 2008; Francí et al. 2006; Polyak and Weinberg 2009). While there has been previously established roles for the local tumor microenvironment or cell-intrinsic factors like mutations for invasive behavior in metastasis (Olmeda et al. 2017; Kaur et al. 2019; Nataraj, Marrocco, and Yarden 2021; Nguyen et al. 2022), more recently, the role of non-genetic, cell-intrinsic factors have been implicated (Arozarena and Wellbrock 2019; Quinn et al. 2021). What is unclear is the role that single cells have in initiating phenotype switching within the primary tumor to begin invasion and dissemination of the tumor. Namely, are the rare cells that are able to leave the primary tumor and invade other tissues intrinsically primed to do so (Quinn et al. 2021), do they leave because of external factors such as their local microenvironment (Kaur et al., 2019; Olmeda et al. 2017), or some combination of the two?

Genetic differences such (i.e., mutations) have often been implicated in driving the transition to a more invasive state underlying metastasis (Nataraj et al., 2021; Nguyen et al., 2022). However, recent work has established the role of non-genetic changes in

regulatory pathways to cause the switch to an invasive phenotype (Arozarena and Wellbrock, 2019; Quinn et al., 2021). For example, this phenotype switching in melanoma is driven by changes in Wnt pathway signaling and by factors in the local tumor milieu.

Here, we show that within clonal melanoma cell lines there are rare and highly invasive subpopulations. Moreover, this phenotype is transient and marked by the expression of *SEMA3C*. The transcription factor *NKX2.2* also negatively regulates the formation of the invasive subpopulation.

3.2 Results

3.2.1 *SEMA3C* marks a rare and invasive population

Invasiveness of cells was measured using polytetrafluoroethylene transwells and a serum gradient to encourage invasion. The so-called early invading cells are the small percentage of cells that invade through the transwells in the first 8 hours of the assay (**Figure 3.1**). To identify a candidate marker of the behavior, early invading cells, late-invading cells, and non-invading cells had their transcriptomes profiled via RNA sequencing. *SEMA3C,* a gene whose protein product is expressed on the cell surface, was identified as a differentially expressed gene that marked early invading cells (**Figure 3.2**). To establish *SEMA3C* as a *bona fide* marker of the early invading population, cells were sorted based on the degree of *SEMA3C* expression using flow-assisted cell sorting (**Figure 3.3**) and their rate of invasiveness was measured using the transwell assay. In fact, SEMA3C-high cells were far more invasive than SEMA3C-low cells and the overall population (**Figure 3.4**), suggesting *SEMA3C* is a marker of the early-invading population.

3.2.2 *NKX2.2* is a transcription factor that promotes the invasive subpopulation

ATAC-sequencing was performed on early-invading, late-invading, and non-invading FS4 cells to characterize differences in chromatin accessibility and identify putative regulatory factors. Overall, a relatively small amount (1107) of differential peaks were identified that characterized the early-invading population. The homeobox-domain containing transcription factor *NKX2.2* (also commonly referred to as *NKX2-2*) was identified as a putative regulator of the early-invading phenotype. To test the role of NKX*2-2* in creating an early invading subpopulation, we knocked it out using CRISPR-Cas9-mediated genome editing. Much to our surprise, rather than making cells less invasive as we hypothesized, knockout of *NKX2-2* caused 1205Lu cells to become more invasive and proliferate at a faster rate (**Figure 3.5**). Principal component analysis of RNA and ATAC-sequencing data demonstrated that cells with an active guide targeting *NKX2-2* separately in principal component space for both data modalities (**Figure 3.6**). While there was some overlap in the transcriptional profile between *NKX2-2* knockout 1205Lu cells and early invading 1205Lu and FS4 cells (**Figures 3.6, 3.7, 3.8, 3.9**), the knockout cells were still transcriptionally distinct. Most of the common differentially expressed genes were those downregulated in early invaders and in *NKX2-2* knockout cells which were genes involved in cell migration, cell motility, and extracellular matrix organization (**Figure 3.10**).

3.3 Discussion

Overall, these findings indicate that clonal melanoma cells can have a cell-intrinsic, non-genetic ability to become invasive. This rare population in our system is characterized by

high expression of the surface protein SEMA3C and enrichment for cells highly expressing SEMA3C enriches for invasive cells. This work also establishes the foundation for *NKX2.2* as a regulator of invasive behavior in melanoma. This is result is particularly interesting given that to date the only mention of *NKX2.2*, which has been implicated in Ewing Sarcoma, and melanoma is an immunohistochemical study mentioning that 2/6 melanoma samples stained positive for *NKX2.2 (Yoshida et al. 2012)*. Future work should better delineate the mechanism by which *NKX2.2* increases both invasiveness and proliferation and look to recapitulate these results *in vivo*.

3.4 Contributions

This chapter contains quotes and figures from Kaur *et al*. published in 2022 in BioRxiv (Kaur et al. 2022). AK, designed, performed and analyzed all experiments. KK curated and performed all analysis on sequencing data generated by AK. Figures 3.1, 3.3, 3.4, and 3.5 were made by AK, while the rest of the figures and associated analyses mentioned in this chapter were done by KK.

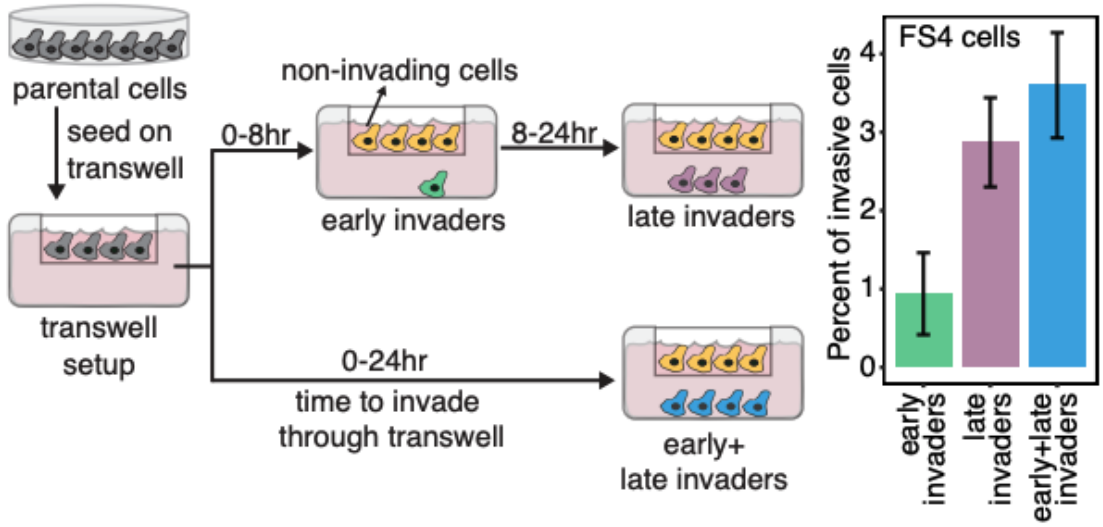

Figure 3.1 A rare, early invading subpopulation of cells is prime for invasion

Schematic showing the transwell assay with definitions of different invasive cell populations and their relative proportions of the total population for the FS4 cell line.

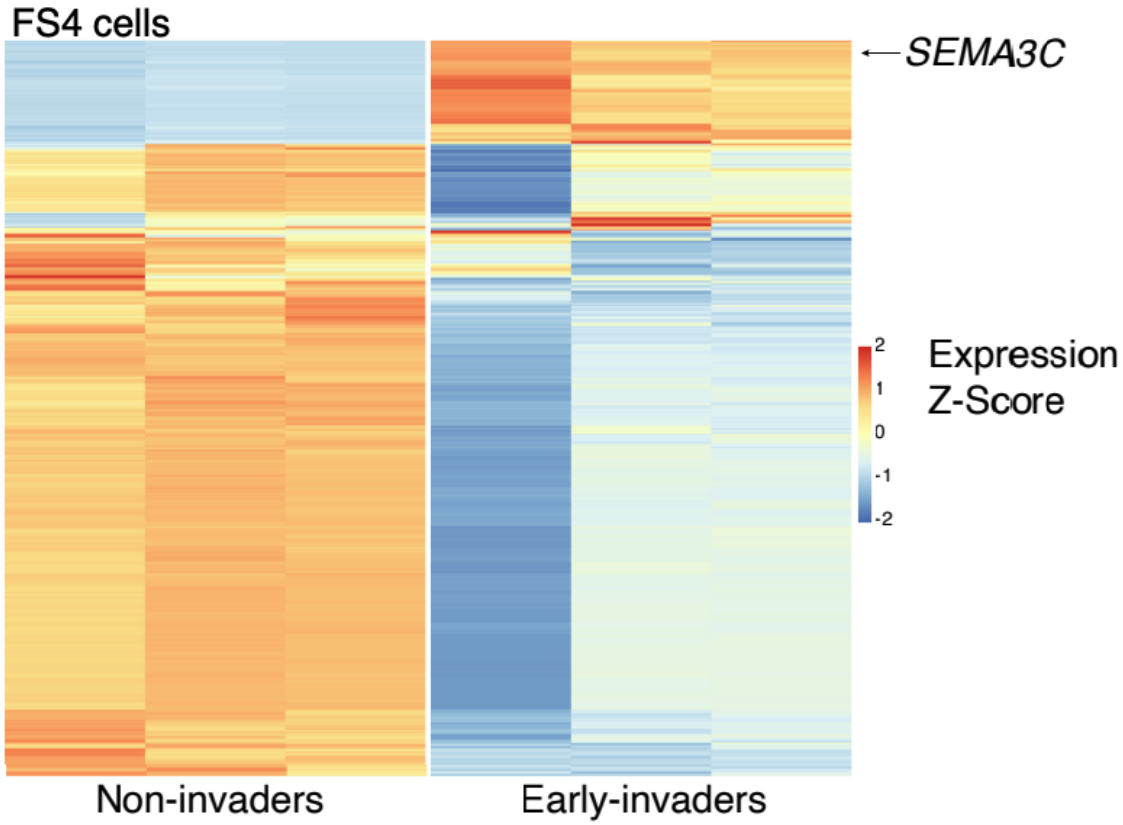

Figure 3.2 RNA-sequencing establishes *SEMA3C* as a potential marker of the early-invading population

Heatmap showing all differentially expressed genes (including *SEMA3C*) between early-invading and non-invading FS4 melanoma cells.

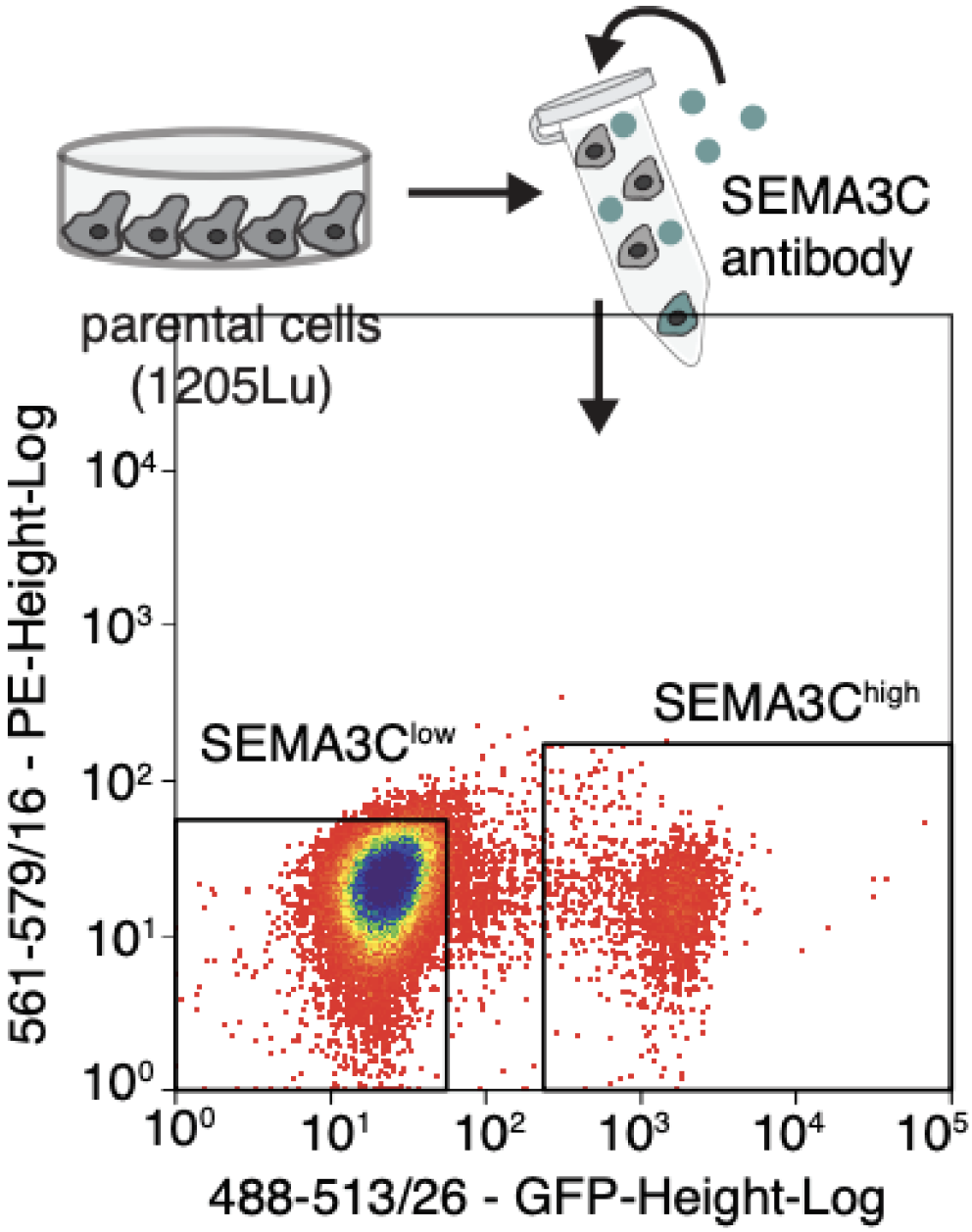

Figure 3.3 Fluorescence-activated cell sorting (FACS) of a 1205Lu melanoma cells based on SEMA3C protein expression

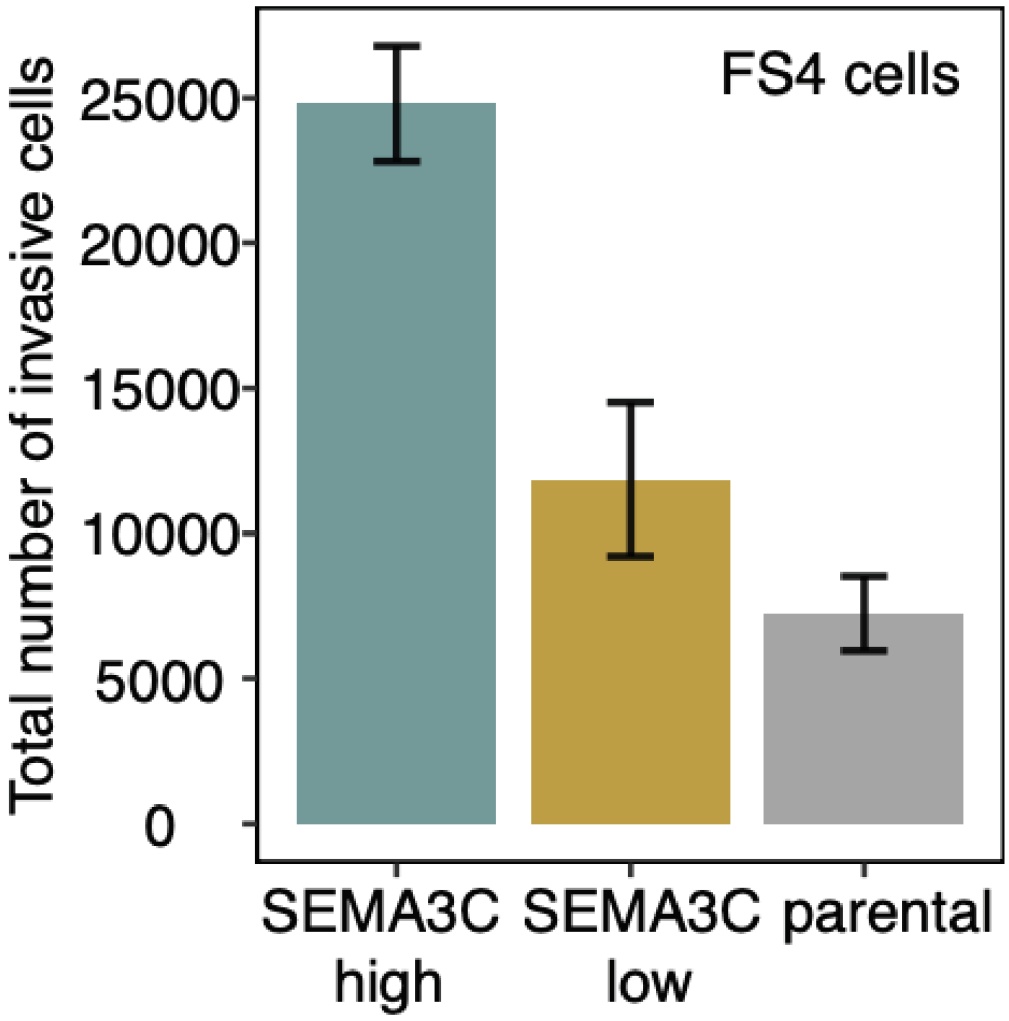

Figure 3.4 FS4 cells highly expressing SEMA3C are more invasive

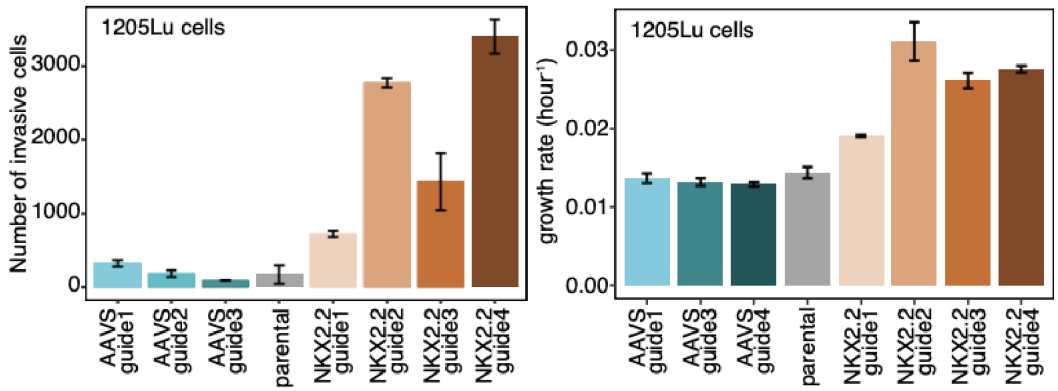

Figure 3.5 *NKX2.2* negatively regulates both invasive and proliferative behaviors in 1205Lu melanoma cells

1205lu melanoma cells expressing either AAVS or NKX2.2 knockout were seeded on the transwell and the number of invading cells was calculated (left). 1205lu melanoma cells expressing either AAVS or NKX2.2 knockout were seeded in tissue culture plates and cells were allowed to grow for 10 days. Cells were imaged every 24 hours and cell counts at different times were determined and used to calculate growth rate of the cells (right). Error bars represent standard error across 3 replicates.

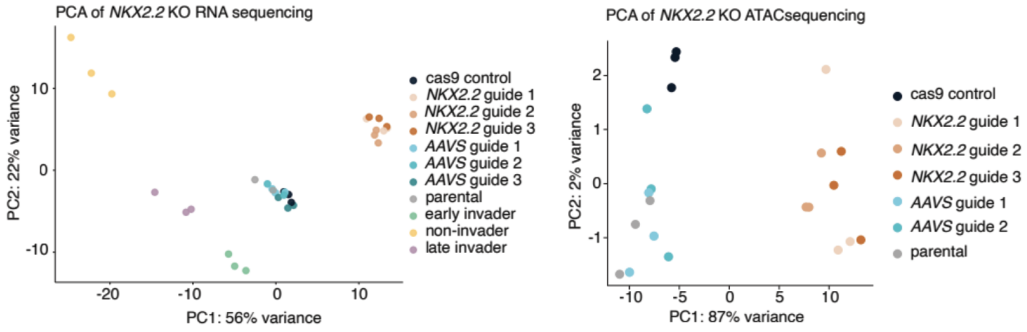

Figure 3.6 *NKX2.2* knockout cells cluster separately in principal component space using gene expression and chromatin accessibility data

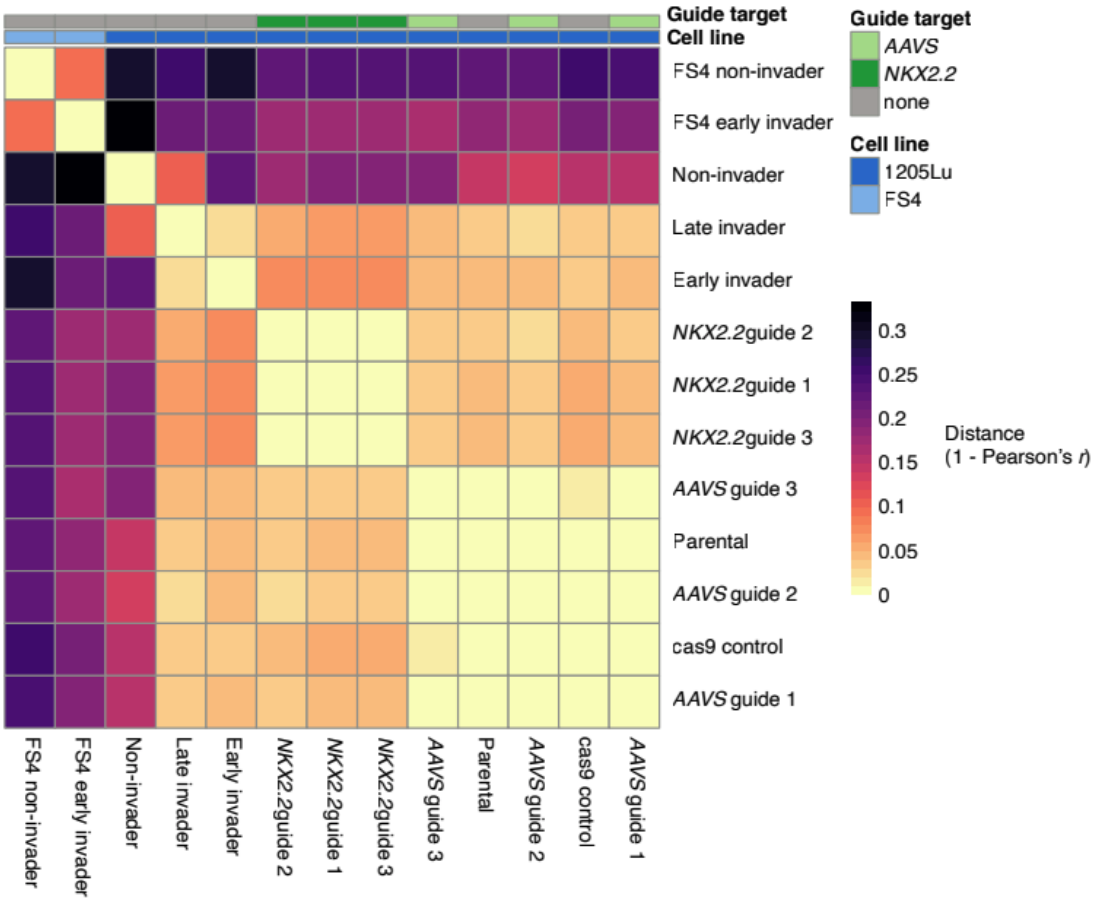

Figure 3.7 Pairwise distance metrics show that *NKX2.2* knockout cells are most similar to 1205Lu early and late invaders.

Matrix of distance (measured by 1 - Pearson's r) between samples' expression for the union of differentially expressed genes between both cell lines

when comparing early invaders to non-invaders.

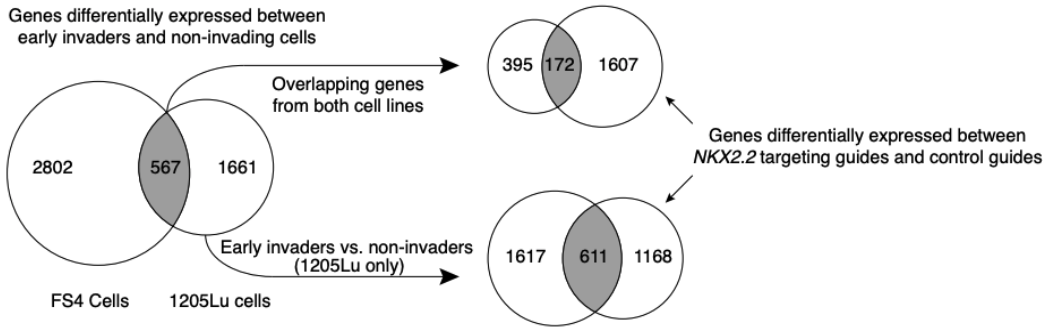

Figure 3.8 Differentially expressed genes show some overlap between cell lines and early-invading cells and *NKX2.2* knockout cells

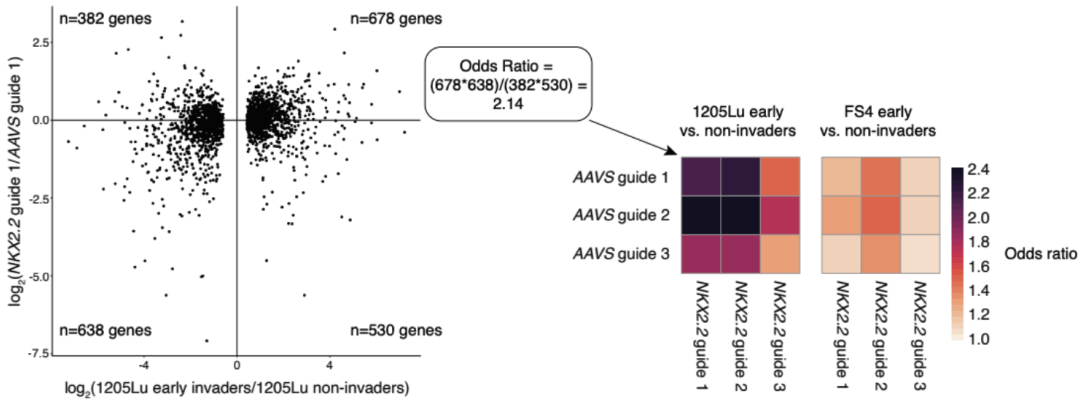

Figure 3.9 Odds ratio analysis looking at similarity of *NKX2.2* knockout cells to 1205Lu and FS4 early invaders

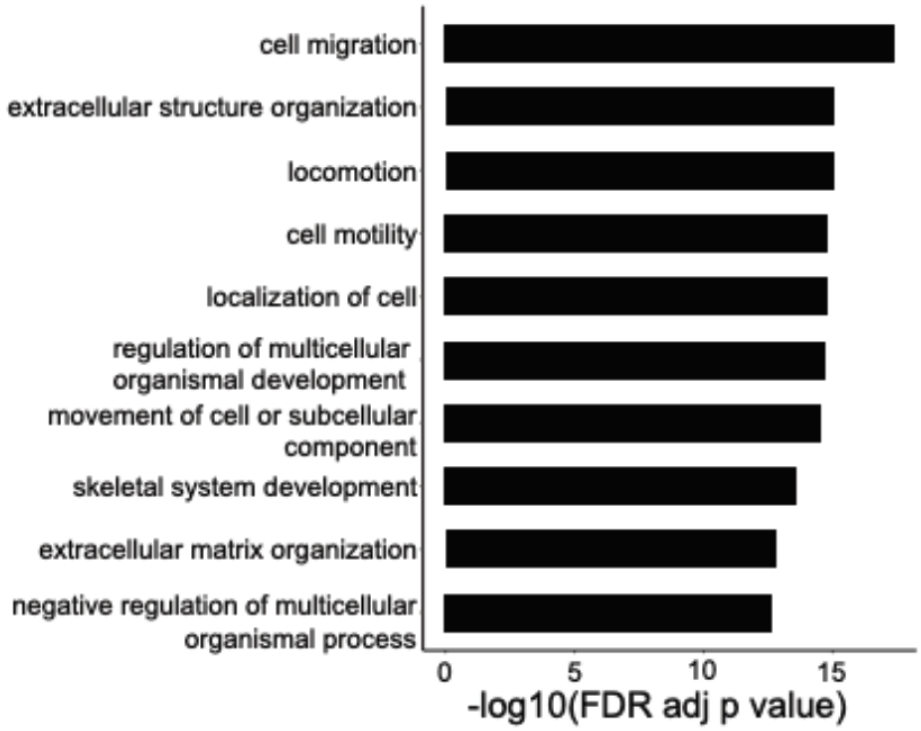

Figure 3.10 Overrepresentation analysis of genes differentially downregulated after NKX2.2 knockout.

CHAPTER 4: CONCLUSIONS AND FUTURE DIRECTIONS

In the work presented in this thesis through two different projects, I both used existing bioinformatic approaches and developed computational techniques to better understand principles of gene regulation both more broadly in the context of response to single-factor perturbations as well as in the context of distant metastasis in melanoma. A systematic analysis of tandem, bulk RNA-seq and ATAC-seq demonstrated that the changes in gene expression from signals such as retinoic acid or TGF-β were not necessarily concordant with changes in chromatin accessibility. Moreover, the genes which had concordant changes between chromatin accessibility and gene expression tended to be less accessible prior to stimulation with signal. Further analysis using multiple types of assigning peaks and subsetting peaks based on computationally predicted transcription factor activity failed to demonstrate any further indication of concordance. These results suggest that at least in the context of these two signals, there are two modes of regulation at play. However, questions remain whether and to what degree these findings hold in other systems and contexts.

Additionally, we identified a highly invasive subpopulation within clonal melanoma cell lines that are marked with the transient, high expression of *SEMA3C*. This subpopulation was shown to drive the distant metastasis, (i.e., those beyond local lymph nodes) in mouse models of melanoma. Using bioinformatic analyses we identified the transcription factor *NKX2.2* as a possible regulator of this invasive state and that its knockout created highly invasive cells. Further studies should better elucidate this factor's role more mechanistically as well as characterize what role it has in metastasis *in vivo*.

4.1 Determining which peaks interact with which genes

Finding which gene(s) a given region of accessibility interacts with is by no means a trivial task. Here, we adopted two different heuristics to map peaks to genes. The first, commonly used both in ATAC-seq and other peak-based genomic profiling methods (Li et al. 2012; Nair et al. 2021) is to simply find the nearest transcription start site and assign the peak to that gene. This creates a one-to-one peak to gene mapping, but since this linear approach fails to take into account the three-dimensional conformation of chromatin and the long range contacts that may occur, we also used a window-based approach. This took every peak within a window around the transcriptional start site creating a many-to-one mapping. However, this method also has its own shortcomings given that certain enhancers, like the sonic hedgehog limb-bud-specific enhancer, can act from over 850 kilobases away (Lettice et al. 2003), much farther away than our largest window of 100 kilobases. Current work involves leveraging deep learning architectures, especially convolutional neural nets to infer from sequence alone *cis*-regulatory elements, predicted transcription factor binding sites, and higher order transcription factor "syntax" (Vaishnav et al. 2022; de Almeida et al. 2022; Novakovsky et al. 2022). While the use of these machine learning-based approaches is still in its infancy, the findings from these studies and further refinements may create better mappings with which to infer concordance.

Alternatively, there are other approaches that can be invoked to further elucidate a given peak's contribution to gene expression change. For example, Weissman and colleagues developed a technique which correlated peaks and the eigenvector of this correlation matrix, named "eigenpeaks", on a gene-by-gene basis. Then, eigenpeaks were correlated with gene expression. This method reduces the covariance of multiple peaks

nearby into one value, and there was rarely more than one eigenpeak per gene, indicating that nearby *cis*-regultory elements typically act in concert in their system (Mold et al. 2022). Moreover, ArchR, a software package developed for single-cell chromatin accessibility analysis tested over 50 models to create a gene score from chromatin accessibility data. The highest scoring models tested on data from bone marrow and peripheral blood monocyte samples used signals from the promoter and gene body and used an exponential decay function to weight the contribution of more distal regulatory elements (Corces et al. 2018). Finally, Cicero, which was also developed for single-cell chromatin accessibility data, uses a combination of co-accessibility correlations with a graphical LASSO and distance penalty to infer a cis-regulatory map (Pliner et al. 2018). While Cicero in theory could be adapted to bulk chromatin accessibility data, it may not be as effective given the model assumes input from many cells rather than the usually limited number of samples done in bulk studies, but its tractability for this purpose should be explored.

Finally, peak-gene pairs can be further informed from topologically associating domains (TADs) measured by Hi-C. However, TAD boundaries are not always informative of gene expression relationships and the disruption of TAD boundaries does not necessarily have an effect on gene expression or development phenotypes (Despang et al. 2019; Ghavi-Helm et al. 2019; Paliou et al. 2019; Williamson et al. 2019; Tena and Santos-Pereira 2021). Furthermore, TAD boundaries are not as stable as previously thought as studies in cancer have demonstrated that reprogramming of binding sites and TAD boundaries is a hallmark of therapy resistance in multiple systems (Achinger-Kawecka et al. 2020; Zhou et al. 2022). A possible “gold-standard” approach to definitively establish accessible peak to gene mappings is to use tandem RNA-seq and

ChIP-seq or HiChIP for the transcription factor(s) of interest to a signaling process along with ATAC-seq. However, the scale and cost of this kind of experiment is significant and can suffer from technical shortcomings of the ChIP-seq/HiChIP techniques in that they are heavily reliant on the affinity of available antibodies for a protein of interest and that each experiment can only interrogate one protein at a time and do not identity interacting proteins that often work in tandem in the regulatory complex. Furthermore, the downstream analysis, like many multi-omics analyses, is by no means straightforward.

4.2 Applying findings related to chromatin accessibility and gene expression to single-cell technologies

The observant reader of this dissertation (and bless you if you have made it this far) will notice that thus far in the discussion of concordance of chromatin accessibility and gene expression measurements, there has been little to no mention of single-cell technologies. Recent advances in barcoding, microfluidics, and robotics have made both scRNA-seq and scATAC-seq as well as both assays in tandem far more straightforward to perform and the results more reproducible. However, as these technologies are still nascent, the sparsity of data produced due to technical dropouts as well as issues with separating biological variation from technical artifacts and batch effects makes it difficult to more systematically analyze concordance between these data (Minnoye et al. 2021). The recent introduction of the Chromium Single Cell Multiome ATAC + gene expression kit (10X Genomics, Pleasanton, CA) has proved to be exciting in that it allows for simultaneous measurements of gene expression and chromatin accessibility in the same cell. However, this method is still somewhat hampered in that gene expression measurements are restricted to nuclear transcripts only.

Despite these limitations, there is reason for optimism that future work looking to determine concordance at a single cell resolution will be more tractable. First, single-cell technologies are improving in their accuracy and precision rapidly as a result of constant innovation. For example, very recent work by Chen and colleagues have established a new method, sequencing of nuclear protein epitope abundance, chromatin accessibility, and the transcriptome in single cells (NEAT-seq). NEAT-seq is able to interrogate all tenets of the central dogma of biology, and as a test case, NEAT-seq was used to identify transcription factors with regulatory activity in creating specific T-cell subsets (A. F. Chen et al. 2022). Furthermore, there is some indication that 'pseudo-bulk' accessibility profiles, i.e. those created from combining the sparse data from all members of the same cluster in low dimensional space can recapitulate accessibility profiles from *bona fide* bulk experiments (Minnoye et al. 2021), but this assumption needs to be more rigorously examined. Another intriguing future direction is to create pseudo-bulk profiles from a multiome experiment and compare the results of the concordance analyses presented in this dissertation to one done in the same system using true bulk sequencing data.

4.3 Further considerations for future work examining concordance

4.3.1 Expanding the palette of transcription factors

Here, we examine the activity of two groups of transcription factors related to the biological signaling of the two perturbations used. As a result, our findings are relevant to the transcriptional effectors of retinoic acid (retinoic acid receptor α and HOX family transcription factors) and TGF-β (SMAD transcription factors) within the context of MCF-7 breast cancer cells being exposed to these signals for approximately three days.

There is no guarantee that other transcription factors would behave similarly within this system, and it remains largely unknown what may happen if one were to interrogate the effects of varying systems, perturbations or time scales. Indeed, as established here and previously, within the context of developmental signals there often is far more concordance between chromatin accessibility changes and gene expression. Along with these findings is the fact that there are lineage defining transcription factors that are able to bind to inaccessible chromatin and make the chromatin accessible for itself and other factors to bind, the so-called pioneer factors (Zaret 2020). Among these are members of the fork head box (FOX) family and GATA family members, and surely the relationship of concordance when examining these factors would be different than our findings.

Alternatively, there is a body of work demonstrating that some factors depend heavily on the pre-existing chromatin accessibility landscape for their binding. For example, the glucocorticoid receptor binds almost exclusively to pre-existing accessible chromatin prior to stimulation with dexamethasone (John et al. 2011), and that activator protein 1 (AP-1) establishes this binding pattern for the glucocorticoid receptor by maintaining chromatin accessibility (Biddie et al. 2011). Similarly, the lineage-defining transcription factor Foxp3 binds to preformed accessible sites established by its structural homolog, Foxo1, to establish regulatory T cell identity (Samstein et al. 2012). Of note, this process of regulatory T cell specification via Foxp3 is considered a 'late differentiation' process, as the precursor cell state, the mature naive CD4+ T cell is considered mature. These studies looked at chromatin accessibility data along with ChIP-seq data of the factor of interest, and further work looking to examine concordance in these contexts should also include transcriptomic data. Thus, there is a need to further examine a variety of factors in a variety of contexts using approaches established in this

dissertation as a starting framework to gain a better understanding which underlying regulatory relationships are more unique to specific contexts and which are more general principles of eukaryotic gene regulation.

Finally, a limitation of the findings in the above work is that they are from a clonal cell line system. Further studies should investigate to what degree our findings are applicable to *in vivo* systems responding to physiologic situations. Indeed, some work has demonstrated similar findings to ours from primary placental tissue samples (Starks et al. 2019). A particularly promising primary system are cells from the peripheral immune compartment as they are often poised to react quickly to signals such as lipopolysaccharide, and primary cells are relatively straightforward to collect and manipulate *ex vivo*. Another often overlooked opportunity due to a bias toward mammalian and yeast systems for examining eukaryotic gene regulation is to examine concordance between chromatin accessibility and gene expression within the plant kingdom. In fact, there exists some precedent of this in the literature, especially using the model system *Arabidopsis thaliana* (Farmer et al. 2021). The wide array of aneuploidy in plant genomes as well as the rich array of physiological process that requires precise transcriptional regulation such as flowering, phototropism, and thigmotropism, to name a few, present an invaluable opportunity to better understand eukaryotic gene regulation.

4.3.2 The issue of timing

In our work, MCF-7 cells were continuously exposed to perturbations for 72 hours and at the end of this time period cells were split into two pots for either chromatin accessibility or gene expression measurements. The underlying assumption of this approach is that 72 hours of continuous exposure is sufficient to induce all changes in gene expression

and chromatin accessibility to measure. Other groups (Hota et al. 2022; Ramirez et al. 2017; Bunina et al. 2020) have instead used serial measurements at multiple time points after a perturbation and looked at the concordance between changes in chromatin accessibility at a given time point and gene expression at a later time point. Whether or not the assumption of ordinality of accessibility change to expression change is correct is subject to debate and may miss secondary changes to chromatin accessibility in response to gene expression changes. However, this more temporally aware approach is nonetheless important, and care should be taken to consider these dynamics when examining concordance in the future.

4.3.3 Disentangling possible confounding effects due to the cell cycle

As many studies, including our own, examine or will examine the concordance between these data in actively cycling cells, it is important to remember that during the process of mitosis, transcription is halted, chromatin condenses into chromosomes in anticipation of metaphase and the resulting progeny must re-establish at least part of the transcriptional program of their antecedents. Cell cycle can indeed be such an important confounder that scRNA-seq analyses routinely regresses out the effect of cell cycle based on transcriptionally inferred cell cycle scores (Nestorowa et al. 2016). Further work should be done to consider what effect, if any, this may have on our findings. A considerable body of work exists on “mitotic bookmarking,” or the retention of specific transcription factors at target loci on mitotic chromosomes such that the necessary transcriptional information can be propagated to progeny (Zaidi et al. 2010; Teves et al. 2016). Hsiung and colleagues used a murine erythroblast model to compare chromatin accessibility between cells undergoing mitosis and those in interphase to demonstrate that chromatin accessibility at the macromolecular level is largely independent of cell

cycle (Hsiung et al. 2015). However, future work has to consider whether or not that is the case for the system of interest as well as whether any transcription factors salient to the biological questions are likely to be retained for bookmarking or evicted during mitosis.

4.3.4 Transcription factor footprinting in chromatin accessibility data

A commonly cited limitation of methods for measuring chromatin accessibility genome-wide is that inferring the specific transcription factor bound to an accessible region from these data is non-trivial. Thus, it is necessary to corroborate findings with further mechanistic studies including, but not limited to, genetic perturbation of transcription factors, ChIP-seq, or measuring of nascent RNA (Minnoye et al. 2021). There are also more technical considerations for transcription factor footprinting. Historically, DNase-seq has continued to outperform ATAC-seq for transcription factor footprinting (Sung, Baek, and Hager 2016), but more recent advances have begun to also better adapt footprinting for ATAC-seq, including better modeling the effects of Tn5 transposase bias (Karabacak Calviello et al. 2019). Regardless of modality, to accurately identify transcription factor binding, libraries produced must be sequenced at a great depth. Thus, with the current state of the art, there are limitations in inferring the activity of transcription factors using accessibility data.

5.3.5 An accessible peak does not a site of transcription make

Another important level of transcriptional regulation not addressed in the contents of Chapter 2 are the host of the post-translational modifications, including methylation, acetylation, phosphorylation, methylation, SUMOylation, among others (Strahl and Allis 2000). By altering the electronic charge of histone tails, these modifications can alter the

binding of histone tails to DNA and therefore gene expression (Kouzarides 2007; Yanjun Zhang et al. 2021).

Of note is histone acetylation, which reduces the positive charge of lysine residues in the histone tail, leaving DNA exposed (Bannister and Kouzarides 2011). Thus, histone acetylation is often considered an active histone mark (Pogo, Allfrey, and Mirsky 1966; Clayton et al. 1993). While many lysine residues can be acetylated, the acetylation of the 27th residue of histone 3 (H3k27ac) is of particular interest because it is often localized at promoters and enhancers of actively transcribed genes (Creyghton et al. 2010; Rada-Iglesias et al. 2011).

H3K27ac histone modifications are recognized by the p300/cyclic AMP response element-binding protein (CBP) activating protein complex. The p300/CBP complex can then relax chromatin structure at promoters through its intrinsic histone acetyltransferase activity as well as recruiting other acetyltransferases (Q. Jin et al. 2011). Bromodomain and extraterminal domain (BET) proteins also recognize the H3k27ac using their bromodomains and act as scaffolds to recruit other transcription factors and RNA polymerase II to modulate gene expression (Josling et al. 2012; Taniguchi 2016; Benton, Fiskus, and Bhalla 2017).

While in many cases the result of increased H3K27 acetylation is more accessible chromatin and increased gene expression, the exact interplay between these so called “epigenetic” marks is far more complicated. Recent work has only begun to interrogate enhancer elements using an activity-by-contact model along with CRISPRi to test enhancer-gene interactions in 30 genes (Fulco et al. 2019). Future work should expand these methodologies to more genes in more context as well as more rigorously build a model of gene regulation by examining changes not only in gene expression and

chromatin accessibility, but also using localization of important transcription factors and histone modifications.

4.5 Single-cell variability in melanoma metastatic potential

4.5.1 Further characterization of *NKX2.2* deficient cells

One of the most stark findings by Kaur and colleagues was that the rare and transient population of highly invasive cells characterized by high *SEMA3C* expression composed the vast majority of melanoma cells that migrated from the primary tumor and metastasized in the lung (Kaur et al. 2022). Furthermore, *NKX2.2* was identified from ATAC-seq data in the FS4 cell line (cf. the 1205 Lu cell line that *in vivo* and CRISPR-Cas9 knockout studies were done) for characterizing early invading cells. The initial hypothesis was that knock down of this factor whose motif was overrepresented in peaks differentially accessible in early invading cells would lead to loss of the invasive phenotype. Much to our surprise, it not only increased invasiveness, but it also starkly increased the rate of cell proliferation. This finding is notable given that previous literature in melanoma had demonstrated a trade-off or anti-correlation between invasiveness and proliferation, meaning an increase in one attribute usually comes at the price of a decrease in another (Hoek et al. 2006, 2008). Indeed this tradeoff paradigm has been adopted from pareto optimality theory in economics (Debreu 1954) and been applied to biology and division of cellular tasks (Riolo et al. 2013; Hart et al. 2015). Further transcriptional profiling of *NKX2.2* knockout melanoma cells may lend insights into a possible edge case where the rules of pareto optimality may fail. Furthermore, while proliferation and invasiveness increased as a result of knockout using functional assays *in vitro*, an important and logical next step is to adopt an experimental schema

similar to the previously mentioned one which Kaur and colleagues used to demonstrate the invasive potential of cells highly expressing *SEMA3C in vivo*. It is important to recapitulate similar results using a mixture of *NKx2.2* deficient and control cells in murine models to further establish the role of *NKX2.2* as a *bona fide* regulator of this rare, invasive state.

4.5.2 Elucidating the role and mechanism of *NKX2.2*

Our work used a systems biology approach to identify *NKX2.2* as a regulator of the early invading phenotype. However, further work is necessary to more mechanistically characterize what role, if any, *NKX2.2* has in creating this invasive and proliferative phenotype and understanding its relevance to disease pathogenesis. It is promising that there is considerable overlap in genes downregulated in NKX2.2 knockout cells compared to safe harbor controls and those downregulated early invaders versus non-invaders. This seems to indicate that at least in this aspect, the downregulation of a similar set of genes related to cell migration and the extracellular matrix have their expression modulated by *NKX2.2*. As a homeobox domain-containing protein, *NKX2.2* is most commonly implicated in the morphogenesis of the central nervous system (Lovrics et al. 2014) and pancreatic beta cell function (Raum et al. 2006), but there is a dearth of literature on the transcription factors role in cancer, especially in the context of melanoma. Interestingly, while *NKX2.2* is implicated as necessary for oncogenic transformation in Ewing's sarcoma with an EWS/FLI fusion (Smith et al. 2006), the only mention of *NKX2.2* and melanoma in the literature to the best of my knowledge is a paper demonstrating *NKX2.2* as a useful immunohistochemical marker of Ewing sarcoma. This study looked at other small round cell tumors and noted that 2/6 of malignant melanomas tested also stained positive for *NKX2.2 (Yoshida et al. 2012)*,

indicating that as far as *NKX2.2* and melanoma are concerned, we are in *terra incognita*. A useful first step is to more definitively look at binding of *NKX2.2* in both bulk and highly invasive populations using ChIP-seq or HiChIP. While these techniques are by no means trivial, they would identify binding and long-range interaction patterns in these cells to begin to establish the *NKX2.2* regulome specifically in melanoma and melanoma metastatic potential.

4.6 Concluding remarks

Throughout the course of this dissertation, I have demonstrated through systematic analysis of chromatin accessibility and gene expression data that there are two distinct groups of gene expression changes in response to single-factor perturbations: those with concordant accessibility changes and those without. The proposed future experiments would explore how these results hold in other systems or for other transcription factors. Finally, I have begun to lay the foundation of the role of the transcription factor *NKX2.2* in invasive behavior in melanoma metastatic melanoma. Taken together, these findings will not only help better delineate the fundamental regulatory axioms at play in transcriptional regulation, but also help deliver insights to help inform stem cell-based therapeutics and more effective cancer therapies.

CHAPTER 5: MATERIALS AND METHODS

5.1 PCA of RNA and ATAC-sequencing samples

Principal component analysis and visualization of RNA-seq and ATAC-seq samples was performed using raw counts and performing a variance stabilizing transform. Results were visualized using functions from the R DESeq2 package (Love, Huber, and Anders 2014).

5.2 Bulk RNA-sequencing analysis

Initial RNA sequencing analysis was performed as previously (Goyal et al. 2021). Briefly, reads were aligned to the hg38 assembly using STAR v.2.7.1a and counted uniquely mapped reads with HTSeq v0.6.1 and hg38 GTF file from Ensembl (release 90). We used DESeq2 v1.22.2 in R 3.5.1 using a minimum absolute-value log-fold-change of 0.5 and a q value of 0.05. For genes with multiple annotated transcriptional start sites, we used the 'canonical' transcription start site from the knownCanonical table from GENCODE v29 in the UCSC Table Browser.

We performed functional over-representation and gene set enrichment analysis (Subramanian et al, 2005) of upregulated transcripts in the high dose retinoic acid and high dose TGF-β using clusterProfiler v4.0.5 and enrichplot v1.12.3 (T. Wu et al. 2021). P values for the over-representation analysis were adjusted using a false discovery rate approach. We used the C5 ontology and H hallmark curated gene sets from the Molecular Signatures Database (MSigDB) v7.4 (Liberzon et al. 2011, 2015) as reference gene sets to compare our upregulated genes to.

5.3 ATAC-sequencing analysis

ATAC-seq alignment and peak calling was performed as previously (Sanford et al, 2020). We aligned peaks to the hg38 assembly using bowtie2 v2.3.4.1, and filtered out low-quality alignments with samtools v1.96, removed duplicate read pairs with picard 1.96, and used custom Python scripts along with bedtools v2.25.0 to create alignment files with inferred Tn5 insertion points. We called peaks using MACS2 (Zhang et al. 2008) v2.1.1.20160309 with the command, 'macs2 callpeak --nomodel --nolambda --keep-dup all --call-summits -B --SPMR --format BED -q 0.05 --shift 75 --extsize 150'.

Since we had three biological replicates per condition, we used a majority rule approach to retain only summits that were found in at least two replicates (Yang et al. 2014). Using these condition-specific peak files, we used bedtools to create a consensus peak file by merging each individual condition's peak summit file together in a manner that disallowed overlapping peaks. We used bedtools merge command 'bedtools merge -d 50' to combine features within 50 base pairs of each other into a single peak after testing multiple merge distances. We used the number of ATAC-seq fragment counts at each peak in this merged consensus peak file for differential peak analysis.

We used the custom peak analysis algorithm from Sanford et al., 2020 that took advantage of additional ethanol control conditions to estimate false discovery rate in ethanol controls to then identify differential peaks. Briefly, reads were quantified for each peak in the master consensus file and fragments at each peak were normalized to correct for differences in total sequencing depth using the equation:

sample's total reads in peaks/mean number of reads in peaks across all samples. Next, an estimated false discovery rate was calculated in each cell of a 50x50 grid containing 50 exponentially-spaced steps of minimum fold-change values (ranging from 1.5-10) and

50 exponentially-spaced steps of minimum number of normalized fragment counts in the condition with the greater number of counts (ranging from 30 to 237 or 10 to 237). The estimated false discovery rate (FDR) was calculated using the equation: estimated FDR = (no. of conditions)(est. number of false positive peaks per condition)total number of differential peaks in experimental conditions. After calculating the estimated FDR in each cell of the 50x50 grid, we then pooled together differential peaks contained in any cell with an FDR less than 0.25%.

We performed motif analysis on our set of differential peaks using chromVAR v1.8.0 (Schep et al. 2017), its associated cisBP database of transcription factor motifs, and the motifmatchR package from bioconductor. To decrease the variance of the transcription factor motif deviations scores, we pooled together the different dosages of retinoic acid or TGF-β. The chromVAR code was modified to extract an internal metric that equals the fractional change in fragment counts at motif-containing peaks for a given motif.

5.4 Hematopoietic differentiation data processing

We used preexisting RNA- and DNase I-seq data (aligned to genome assembly hg19) of hematopoietic differentiation (González, Setty, and Leslie 2015) to compare against our data. We used data from the website provided in the paper (http://cbio.mskcc.org/public/Leslie/Early_enhancer_establishment/) to download annotations of peaks (peaksTable.txt), counts of DNase-seq (DNaseCnts.txt), and RNA-seq counts (RNAseqCnts.txt). Counts presented in these data files were quantile normalized and averaged when biological replicates were available. We filtered peaks with “CD14” or “CD34” under the “accessPattern” annotation to choose for peaks that were relevant for comparing HSPCs to monocytes. We used a log2 fold change of greater

than or equal to 2 as a cutoff for assigning differential peaks. We used the preexisting annotations of genes for each peak for peak-gene mapping. For determining the log2 fold change in gene expression we discarded genes whose maximum expression value across the two conditions was fewer than 5 quantile-normalized units.

For visualization of this data set with our own accessibility data, we realigned raw fastq files DNase-seq files to the hg38 assembly using bowtie v2.3.4.1 and filtered out low-quality alignments with samtools v1.1 to generate new .bam alignment files. The alignment files were combined using samtools merge in a single .bam file per cell type. Bam files were converted to .bigWig format using deeptools 3.5.1 (Ramírez et al, 2016) "bamCoverage -- normalizeUsing CPM" to create a 'consensus' .bigWig for visualization. Peaks for CD34+ and CD14+ samples were made by filtering peaks annotated for these populations in the "accessPattern" column and creating separate .bed files using a custom script. The peak location in these .bed files were then lifted over from hg19 to hg38 using UCSC hgLiftOver. For comparing the overlap of peaks between data sets, we created consensus peak sets across all sample types and used the bedtools intersect function to quantify the proportion of peaks that intersected between the hematopoietic differentiation and signaling data.

5.5 Peak annotation

Peaks were annotated using ChIPseeker (Yu, Wang, and He 2015) to determine the relative proportion of features in the data from González et al., 2015 (DNAse-seq) and Sanford et al., 2020 (ATAC-seq). For ease of visualization, certain categories like the three promoter categories were collapsed into one. ChiPseeker was also used to identify the nearest transcriptional start site to a gene used for the nearest integration approach described below. For making scatter plots of change in accessibility versus change in

expression annotated by peak feature, a custom script was used to combine annotations from ChIPseeker into four categories: downstream, gene body, integenic, and promoter.

For each of the top 150 most variable transcription factor motifs we identified using differential accessibility analysis, we used the R bioconductor motifmatchR package to annotate both the number of motif matches and a log-likelihood match score for each peak.

5.6 RNA-seq and ATAC-seq data integration

We employed two methods for assigning peaks to genes. In the 'nearest' approach, we used annotation from ChIPseeker to assign each peak to the nearest transcriptional start site. With this method, each peak is uniquely mapped to a single gene. In the 'window' approach we used a window of 50 kilobases on either side of the transcriptional start site (100 kilobases in total) to assign peaks to a gene, which could result in a peak being assigned to multiple genes.

5.7 Track Visualization

We visualized accessibility data using the web based version of integrative genomics viewer (IGV) (Robinson et al. 2011, 2020). We prepared accessibility data for visualization by taking consensus files and converting them to .bigWig file format with either fragments per million or counts per million normalization. Bed files for identifying peaks were created using custom scripts.

5.8 Statistics and software

Unless otherwise stated, we performed analyses using R v4.1.0 with data manipulation and visualization done with tidyverse v1.3.1 (Wickham et al. 2019) and ggpubr v0.4.0. We used a Kolmogorov-Smirnov test to compare means. Unless otherwise stated, 95%

confidence intervals for Pearson's r were calculated by bootstrapping using 10,000 replicates.

5.9 Reproducible analyses

All data and remaining code for these analyses can be found at https://www.dropbox.com/sh/qbjuagz511c072g/AAChvYMjdoG7A0eNdqbEmaUla?dl=0. Analyses were done in R or on the command line. We used a selection of color-blind friendly colors from a custom palette.

BIBLIOGRAPHY

Achinger-Kawecka, Joanna, Fatima Valdes-Mora, Phuc-Loi Luu, Katherine A. Giles, C. Elizabeth Caldon, Wenjia Qu, Shalima Nair, et al. 2020. "Epigenetic Reprogramming at Estrogen-Receptor Binding Sites Alters 3D Chromatin Landscape in Endocrine-Resistant Breast Cancer." *Nature Communications* 11 (1): 1–17.

Ackermann, Amanda M., Zhiping Wang, Jonathan Schug, Ali Naji, and Klaus H. Kaestner. 2016. "Integration of ATAC-Seq and RNA-Seq Identifies Human Alpha Cell and Beta Cell Signature Genes." *Molecular Metabolism* 5 (3): 233–44.

"Agave Americana (century Plant)." n.d. CABI. Accessed May 26, 2022. https://www.cabi.org/isc/datasheet/3851.

Alexaki, Vasileia-Ismini, Delphine Javelaud, Leon C. L. Van Kempen, Khalid S. Mohammad, Sylviane Dennler, Flavie Luciani, Keith S. Hoek, et al. 2010. "GLI2-Mediated Melanoma Invasion and Metastasis." *Journal of the National Cancer Institute*.

Almeida, Bernardo P. de, Franziska Reiter, Michaela Pagani, and Alexander Stark. 2022. "DeepSTARR Predicts Enhancer Activity from DNA Sequence and Enables the de Novo Design of Synthetic Enhancers." *Nature Genetics* 54 (5): 613–24.

Ampuja, M., T. Rantapero, A. Rodriguez-Martinez, M. Palmroth, E. L. Alarmo, M. Nykter, and A. Kallioniemi. 2017. "Integrated RNA-Seq and DNase-Seq Analyses Identify Phenotype-Specific BMP4 Signaling in Breast Cancer." *BMC Genomics* 18 (1): 68.

Angelini, Claudia, and Valerio Costa. 2014. "Understanding Gene Regulatory Mechanisms by Integrating ChIP-Seq and RNA-Seq Data: Statistical Solutions to Biological Problems." *Frontiers in Cell and Developmental Biology* 2 (September): 51.

Arozarena, Imanol, and Claudia Wellbrock. 2019. "Phenotype Plasticity as Enabler of Melanoma Progression and Therapy Resistance." *Nature Reviews. Cancer* 19 (7): 377–91.

Åström, Karl Johan, and Richard M. Murray. 2010. *Feedback Systems*. Princeton University Press.

Bannister, Andrew J., and Tony Kouzarides. 2011. "Regulation of Chromatin by Histone Modifications." *Cell Research* 21 (3): 381–95.

Barski, Artem, Suresh Cuddapah, Kairong Cui, Tae-Young Roh, Dustin E. Schones, Zhibin Wang, Gang Wei, Iouri Chepelev, and Keji Zhao. 2007. "High-Resolution Profiling of Histone Methylations in the Human Genome." *Cell* 129 (4): 823–37.

Benton, Christopher B., Warren Fiskus, and Kapil N. Bhalla. 2017. "Targeting Histone Acetylation: Readers and Writers in Leukemia and Cancer." *Cancer Journal* 23 (5): 286–91.

Biddie, Simon C., Sam John, Pete J. Sabo, Robert E. Thurman, Thomas A. Johnson, R. Louis Schiltz, Tina B. Miranda, et al. 2011. "Transcription Factor AP1 Potentiates Chromatin Accessibility and Glucocorticoid Receptor Binding." *Molecular Cell* 43 (1): 145–55.

Boyle, Alan P., Sean Davis, Hennady P. Shulha, Paul Meltzer, Elliott H. Margulies, Zhiping Weng, Terrence S. Furey, and Gregory E. Crawford. 2008. "High-

Resolution Mapping and Characterization of Open Chromatin across the Genome." *Cell* 132 (2): 311–22.
Buenrostro, Jason D., Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. 2013. "Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position." *Nature Methods* 10 (12): 1213–18.
Buenrostro, Jason D., Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. 2015. "Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation." *Nature* 523 (7561): 486–90.
Bunina, Daria, Nade Abazova, Nichole Diaz, Kyung-Min Noh, Jeroen Krijgsveld, and Judith B. Zaugg. 2020. "Genomic Rewiring of SOX2 Chromatin Interaction Network during Differentiation of ESCs to Postmitotic Neurons." *Cell Systems* 10 (6): 480–94.e8.
Cao, Yi, Zizhen Yao, Deepayan Sarkar, Michael Lawrence, Gilson J. Sanchez, Maura H. Parker, Kyle L. MacQuarrie, et al. 2010. "Genome-Wide MyoD Binding in Skeletal Muscle Cells: A Potential for Broad Cellular Reprogramming." *Developmental Cell* 18 (4): 662–74.
Chen, Amy F., Benjamin Parks, Arwa S. Kathiria, Benjamin Ober-Reynolds, Jorg J. Goronzy, and William J. Greenleaf. 2022. "NEAT-Seq: Simultaneous Profiling of Intra-Nuclear Proteins, Chromatin Accessibility and Gene Expression in Single Cells." *Nature Methods* 19 (5): 547–53.
Cheng, Chao, Roger Alexander, Renqiang Min, Jing Leng, Kevin Y. Yip, Joel Rozowsky, Koon-Kiu Yan, et al. 2012. "Understanding Transcriptional Regulation by Integrative Analysis of Transcription Factor Binding Data." *Genome Research* 22 (9): 1658–67.
Cheng, Chao, and Mark Gerstein. 2012. "Modeling the Relative Relationship of Transcription Factor Binding and Histone Modifications to Gene Expression Levels in Mouse Embryonic Stem Cells." *Nucleic Acids Research* 40 (2): 553–68.
Cheng, Chao, Koon-Kiu Yan, Woochang Hwang, Jiang Qian, Nitin Bhardwaj, Joel Rozowsky, Zhi John Lu, et al. 2011. "Construction and Analysis of an Integrated Regulatory Network Derived from High-Throughput Sequencing Data." *PLoS Computational Biology* 7 (11): e1002190.
Cheng, Chao, Koon-Kiu Yan, Kevin Y. Yip, Joel Rozowsky, Roger Alexander, Chong Shou, and Mark Gerstein. 2011. "A Statistical Framework for Modeling Gene Expression Using Chromatin Features and Application to modENCODE Datasets." *Genome Biology* 12 (2): R15.
Chen, Xi, Ricardo J. Miragaia, Kedar Nath Natarajan, and Sarah A. Teichmann. 2018. "A Rapid and Robust Method for Single Cell Chromatin Accessibility Profiling." *Nature Communications* 9 (1): 5345.
Clayton, A. L., T. R. Hebbes, A. W. Thorne, and C. Crane-Robinson. 1993. "Histone Acetylation and Gene Induction in Human Cells." *FEBS Letters* 336 (1): 23–26.
Corces, M. Ryan, Jason D. Buenrostro, Beijing Wu, Peyton G. Greenside, Steven M. Chan, Julie L. Koenig, Michael P. Snyder, et al. 2016. "Lineage-Specific and Single-Cell Chromatin Accessibility Charts Human Hematopoiesis and Leukemia Evolution." *Nature Genetics* 48 (10): 1193–1203.
Corces, M. Ryan, Jeffrey M. Granja, Shadi Shams, Bryan H. Louie, Jose A. Seoane, Wanding Zhou, Tiago C. Silva, et al. 2018. "The Chromatin Accessibility

Landscape of Primary Human Cancers." *Science* 362 (6413). https://doi.org/10.1126/science.aav1898.
Corces, M. Ryan, Alexandro E. Trevino, Emily G. Hamilton, Peyton G. Greenside, Nicholas A. Sinnott-Armstrong, Sam Vesuna, Ansuman T. Satpathy, et al. 2017. "An Improved ATAC-Seq Protocol Reduces Background and Enables Interrogation of Frozen Tissues." *Nature Methods* 14 (10): 959–62.
Creyghton, Menno P., Albert W. Cheng, G. Grant Welstead, Tristan Kooistra, Bryce W. Carey, Eveline J. Steine, Jacob Hanna, et al. 2010. "Histone H3K27ac Separates Active from Poised Enhancers and Predicts Developmental State." *Proceedings of the National Academy of Sciences of the United States of America* 107 (50): 21931–36.
Cusanovich, Darren A., Riza Daza, Andrew Adey, Hannah A. Pliner, Lena Christiansen, Kevin L. Gunderson, Frank J. Steemers, Cole Trapnell, and Jay Shendure. 2015. "Multiplex Single Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing." *Science* 348 (6237): 910–14.
Debreu, G. 1954. "VALUATION EQUILIBRIUM AND PARETO OPTIMUM." *Proceedings of the National Academy of Sciences of the United States of America* 40 (7): 588–92.
Despang, Alexandra, Robert Schöpflin, Martin Franke, Salaheddine Ali, Ivana Jerković, Christina Paliou, Wing-Lee Chan, et al. 2019. "Functional Dissection of the Sox9-Kcnj2 Locus Identifies Nonessential and Instructive Roles of TAD Architecture." *Nature Genetics* 51 (8): 1263–71.
Dong, Xianjun, Melissa C. Greven, Anshul Kundaje, Sarah Djebali, James B. Brown, Chao Cheng, Thomas R. Gingeras, et al. 2012. "Modeling Gene Expression Using Chromatin Features in Various Cellular Contexts." *Genome Biology* 13 (9): R53.
Emert, Benjamin L., Christopher J. Cote, Eduardo A. Torre, Ian P. Dardani, Connie L. Jiang, Naveen Jain, Sydney M. Shaffer, and Arjun Raj. 2021. "Variability within Rare Cell States Enables Multiple Paths toward Drug Resistance." *Nature Biotechnology*, February. https://doi.org/10.1038/s41587-021-00837-3.
Farmer, Andrew, Sandra Thibivilliers, Kook Hui Ryu, John Schiefelbein, and Marc Libault. 2021. "Single-Nucleus RNA and ATAC Sequencing Reveals the Impact of Chromatin Accessibility on Gene Expression in Arabidopsis Roots at the Single-Cell Level." *Molecular Plant* 14 (3): 372–83.
Feng, Jian, Matthew Wilkinson, Xiaochuan Liu, Immanuel Purushothaman, Deveroux Ferguson, Vincent Vialou, Ian Maze, et al. 2014. "Chronic Cocaine-Regulated Epigenomic Changes in Mouse Nucleus Accumbens." *Genome Biology* 15 (4): R65.
Francí, C., M. Takkunen, N. Dave, F. Alameda, S. Gómez, R. Rodríguez, M. Escrivà, et al. 2006. "Expression of Snail Protein in Tumor-Stroma Interface." *Oncogene* 25 (37): 5134–44.
Fulco, Charles P., Joseph Nasser, Thouis R. Jones, Glen Munson, Drew T. Bergman, Vidya Subramanian, Sharon R. Grossman, et al. 2019. "Activity-by-Contact Model of Enhancer-Promoter Regulation from Thousands of CRISPR Perturbations." *Nature Genetics* 51 (12): 1664–69.
Ghavi-Helm, Yad, Aleksander Jankowski, Sascha Meiers, Rebecca R. Viales, Jan O. Korbel, and Eileen E. M. Furlong. 2019. "Highly Rearranged Chromosomes Reveal Uncoupling between Genome Topology and Gene Expression." *Nature Genetics* 51 (8): 1272–82.

Gomez-Cabrero, David, Imad Abugessaisa, Dieter Maier, Andrew Teschendorff, Matthias Merkenschlager, Andreas Gisel, Esteban Ballestar, Erik Bongcam-Rudloff, Ana Conesa, and Jesper Tegnér. 2014. “Data Integration in the Era of Omics: Current and Future Challenges.” *BMC Systems Biology* 8 Suppl 2 (March): I1.

González, Alvaro J., Manu Setty, and Christina S. Leslie. 2015. “Early Enhancer Establishment and Regulatory Locus Complexity Shape Transcriptional Programs in Hematopoietic Differentiation.” *Nature Genetics* 47 (11): 1249–59.

Goyal, Yogesh, Ian P. Dardani, Gianna T. Busch, Benjamin Emert, Dylan Fingerman, Amanpreet Kaur, Naveen Jain, et al. 2021. “Pre-Determined Diversity in Resistant Fates Emerges from Homogenous Cells after Anti-Cancer Drug Treatment.” *bioRxiv*. https://doi.org/10.1101/2021.12.08.471833.

Gupta, Piyush B., Christine M. Fillmore, Guozhi Jiang, Sagi D. Shapira, Kai Tao, Charlotte Kuperwasser, and Eric S. Lander. 2011. “Stochastic State Transitions Give Rise to Phenotypic Equilibrium in Populations of Cancer Cells.” *Cell* 146 (4): 633–44.

Guss, K. A., C. E. Nelson, A. Hudson, M. E. Kraus, and S. B. Carroll. 2001. “Control of a Genetic Regulatory Network by a Selector Gene.” *Science* 292 (5519): 1164–67.

Hart, Yuval, Hila Sheftel, Jean Hausser, Pablo Szekely, Noa Bossel Ben-Moshe, Yael Korem, Avichai Tendler, Avraham E. Mayo, and Uri Alon. 2015. “Inferring Biological Tasks Using Pareto Analysis of High-Dimensional Data.” *Nature Methods* 12 (3): 233–35, 3 p following 235.

Hesselberth, Jay R., Xiaoyu Chen, Zhihong Zhang, Peter J. Sabo, Richard Sandstrom, Alex P. Reynolds, Robert E. Thurman, et al. 2009. “Global Mapping of Protein-DNA Interactions in Vivo by Digital Genomic Footprinting.” *Nature Methods* 6 (4): 283–89.

Hoek, Keith S., Ossia M. Eichhoff, Natalie C. Schlegel, Udo Döbbeling, Nikita Kobert, Leo Schaerer, Silvio Hemmi, and Reinhard Dummer. 2008. “In Vivo Switching of Human Melanoma Cells between Proliferative and Invasive States.” *Cancer Research* 68 (3): 650–56.

Hoek, Keith S., Natalie C. Schlegel, Patricia Brafford, Antje Sucker, Selma Ugurel, Rajiv Kumar, Barbara L. Weber, et al. 2006. “Metastatic Potential of Melanomas Defined by Specific Gene Expression Profiles with No BRAF Signature.” *Pigment Cell Research / Sponsored by the European Society for Pigment Cell Research and the International Pigment Cell Society* 19 (4): 290–302.

Hota, Swetansu K., Kavitha S. Rao, Andrew P. Blair, Ali Khalilimeybodi, Kevin M. Hu, Reuben Thomas, Kevin So, et al. 2022. “Brahma Safeguards Canalization of Cardiac Mesoderm Differentiation.” *Nature* 602 (7895): 129–34.

Hsiung, Chris C-S, Christapher S. Morrissey, Maheshi Udugama, Christopher L. Frank, Cheryl A. Keller, Songjoon Baek, Belinda Giardine, et al. 2015. “Genome Accessibility Is Widely Preserved and Locally Modulated during Mitosis.” *Genome Research* 25 (2): 213–25.

Jimenez-Sanchez, G., B. Childs, and D. Valle. 2001. “Human Disease Genes.” *Nature* 409 (6822): 853–55.

Jin, Qihuang, Li-Rong Yu, Lifeng Wang, Zhijing Zhang, Lawryn H. Kasper, Ji-Eun Lee, Chaochen Wang, Paul K. Brindle, Sharon Y. R. Dent, and Kai Ge. 2011. “Distinct Roles of GCN5/PCAF-Mediated H3K9ac and CBP/p300-Mediated H3K18/27ac in Nuclear Receptor Transactivation.” *The EMBO Journal* 30 (2): 249–62.

Jin, Wenfei, Qingsong Tang, Mimi Wan, Kairong Cui, Yi Zhang, Gang Ren, Bing Ni, et al. 2015. “Genome-Wide Detection of DNase I Hypersensitive Sites in Single Cells and FFPE Tissue Samples.” *Nature* 528 (7580): 142–46.

John, Sam, Peter J. Sabo, Robert E. Thurman, Myong-Hee Sung, Simon C. Biddie, Thomas A. Johnson, Gordon L. Hager, and John A. Stamatoyannopoulos. 2011. “Chromatin Accessibility Pre-Determines Glucocorticoid Receptor Binding Patterns.” *Nature Genetics* 43 (3): 264–68.

Josling, Gabrielle A., Shamista A. Selvarajah, Michaela Petter, and Michael F. Duffy. 2012. “The Role of Bromodomain Proteins in Regulating Gene Expression.” *Genes* 3 (2): 320–43.

Karabacak Calviello, Aslıhan, Antje Hirsekorn, Ricardo Wurmus, Dilmurat Yusuf, and Uwe Ohler. 2019. “Reproducible Inference of Transcription Factor Footprints in ATAC-Seq and DNase-Seq Datasets Using Protocol-Specific Bias Modeling.” *Genome Biology* 20 (1): 42.

Kaur, Amanpreet, Brett L. Ecker, Stephen M. Douglass, Curtis H. Kugel 3rd, Marie R. Webster, Filipe V. Almeida, Rajasekharan Somasundaram, et al. 2019. “Remodeling of the Collagen Matrix in Aging Skin Promotes Melanoma Metastasis and Affects Immune Cell Motility.” *Cancer Discovery* 9 (1): 64–81.

Kaur, Amanpreet, Karun Kiani, Dylan Fingerman, Margaret C. Dunagin, Jingxin Li, Ian Dardani, Eric M. Sanford, et al. 2022. “Metastatic Potential in Clonal Melanoma Cells Is Driven by a Rare, Early-Invading Subpopulation.” *bioRxiv*. https://doi.org/10.1101/2022.04.17.488591.

Kelly, Theresa K., Yaping Liu, Fides D. Lay, Gangning Liang, Benjamin P. Berman, and Peter A. Jones. 2012. “Genome-Wide Mapping of Nucleosome Positioning and DNA Methylation within Individual DNA Molecules.” *Genome Research* 22 (12): 2497–2506.

Kiani, Karun, Eric M. Sanford, Yogesh Goyal, and Arjun Raj. 2022. “Changes in Chromatin Accessibility Are Not Concordant with Transcriptional Changes for Single-Factor Perturbations.” *bioRxiv*. https://doi.org/10.1101/2022.02.03.478981.

Klein, Hans-Ulrich, Martin Schäfer, Bo T. Porse, Marie S. Hasemann, Katja Ickstadt, and Martin Dugas. 2014. “Integrative Analysis of Histone ChIP-Seq and Transcription Data Using Bayesian Mixture Models.” *Bioinformatics* 30 (8): 1154–62.

Kouzarides, Tony. 2007. “Chromatin Modifications and Their Function.” *Cell* 128 (4): 693–705.

Lai, Binbin, Weiwu Gao, Kairong Cui, Wanli Xie, Qingsong Tang, Wenfei Jin, Gangqing Hu, Bing Ni, and Keji Zhao. 2018. “Principles of Nucleosome Organization Revealed by Single-Cell Micrococcal Nuclease Sequencing.” *Nature* 562 (7726): 281–85.

Lareau, Caleb A., Fabiana M. Duarte, Jennifer G. Chew, Vinay K. Kartha, Zach D. Burkett, Andrew S. Kohlway, Dmitry Pokholok, et al. 2019. “Droplet-Based Combinatorial Indexing for Massive-Scale Single-Cell Chromatin Accessibility.” *Nature Biotechnology* 37 (8): 916–24.

Lee, Cheol-Koo, Yoichiro Shibata, Bhargavi Rao, Brian D. Strahl, and Jason D. Lieb. 2004. “Evidence for Nucleosome Depletion at Active Regulatory Regions Genome-Wide.” *Nature Genetics* 36 (8): 900–905.

Lee, Hsiu-Hsiang, and Manfred Frasch. 2005. “Nuclear Integration of Positive Dpp Signals, Antagonistic Wg Inputs and Mesodermal Competence Factors during Drosophila Visceral Mesoderm Induction.” *Development* 132 (6): 1429–42.
Lettice, Laura A., Simon J. H. Heaney, Lorna A. Purdie, Li Li, Philippe de Beer, Ben A. Oostra, Debbie Goode, Greg Elgar, Robert E. Hill, and Esther de Graaff. 2003. “A Long-Range Shh Enhancer Regulates Expression in the Developing Limb and Fin and Is Associated with Preaxial Polydactyly.” *Human Molecular Genetics* 12 (14): 1725–35.
Liberzon, Arthur, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. 2015. “The Molecular Signatures Database (MSigDB) Hallmark Gene Set Collection.” *Cell Systems* 1 (6): 417–25.
Liberzon, Arthur, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. 2011. “Molecular Signatures Database (MSigDB) 3.0.” *Bioinformatics* 27 (12): 1739–40.
Li, Guoliang, Xiaoan Ruan, Raymond K. Auerbach, Kuljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Huay Mei Poh, et al. 2012. “Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation.” *Cell* 148 (1-2): 84–98.
Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.
Lovrics, Anna, Yu Gao, Bianka Juhász, István Bock, Helen M. Byrne, András Dinnyés, and Krisztián A. Kovács. 2014. “Boolean Modelling Reveals New Regulatory Connections between Transcription Factors Orchestrating the Development of the Ventral Spinal Cord.” *PloS One* 9 (11): e111430.
Mani, Sendurai A., Wenjun Guo, Mai-Jing Liao, Elinor Ng Eaton, Ayyakkannu Ayyanan, Alicia Y. Zhou, Mary Brooks, et al. 2008. “The Epithelial-Mesenchymal Transition Generates Cells with Properties of Stem Cells.” *Cell* 133 (4): 704–15.
Ma, Shaoqian, and Yongyou Zhang. 2020. “Profiling Chromatin Regulatory Landscape: Insights into the Development of ChIP-Seq and ATAC-Seq.” *Molecular Biomedicine (Online)* 1 (1): 9.
Metri, Rahul, Abhilash Mohan, Jérémie Nsengimana, Joanna Pozniak, Carmen Molina-Paris, Julia Newton-Bishop, David Bishop, and Nagasuma Chandra. 2017. “Identification of a Gene Signature for Discriminating Metastatic from Primary Melanoma Using a Molecular Interaction Network Approach.” *Scientific Reports* 7 (1): 17314.
Mezger, Anja, Sandy Klemm, Ishminder Mann, Kara Brower, Alain Mir, Magnolia Bostick, Andrew Farmer, Polly Fordyce, Sten Linnarsson, and William Greenleaf. 2018. “High-Throughput Chromatin Accessibility Profiling at Single-Cell Resolution.” *Nature Communications* 9 (1): 3647.
Minnoye, Liesbeth, Georgi K. Marinov, Thomas Krausgruber, Lixia Pan, Alexandre P. Marand, Stefano Secchia, William J. Greenleaf, et al. 2021. “Chromatin Accessibility Profiling Methods.” *Nature Reviews Methods Primers* 1 (1): 1–24.
Mittal, Vivek. 2018. “Epithelial Mesenchymal Transition in Tumor Metastasis.” *Annual Review of Pathology* 13 (January): 395–412.
Mold, Jeff E., Martin H. Weissman, Michael Ratz, Michael Hagemann-Jensen, Joanna Hård, Carl-Johan Eriksson, Hosein Toosi, et al. 2022. “Clonally Heritable Gene

Expression Imparts a Layer of Diversity within Cell Types." *bioRxiv*. https://doi.org/10.1101/2022.02.14.480352.
Nair, Venugopalan D., Mital Vasoya, Vishnu Nair, Gregory R. Smith, Hanna Pincas, Yongchao Ge, Collin M. Douglas, Karyn A. Esser, and Stuart C. Sealfon. 2021. "Differential Analysis of Chromatin Accessibility and Gene Expression Profiles Identifies Cis-Regulatory Elements in Rat Adipose and Muscle." *Genomics* 113 (6): 3827–41.
Nataraj, Nishanth Belugali, Ilaria Marrocco, and Yosef Yarden. 2021. "Roles for Growth Factors and Mutations in Metastatic Dissemination." *Biochemical Society Transactions* 49 (3): 1409–23.
Nestorowa, Sonia, Fiona K. Hamey, Blanca Pijuan Sala, Evangelia Diamanti, Mairi Shepherd, Elisa Laurenti, Nicola K. Wilson, David G. Kent, and Berthold Göttgens. 2016. "A Single-Cell Resolution Map of Mouse Hematopoietic Stem and Progenitor Cell Differentiation." *Blood* 128 (8): e20–31.
Nguyen, Bastien, Christopher Fong, Anisha Luthra, Shaleigh A. Smith, Renzo G. DiNatale, Subhiksha Nandakumar, Henry Walch, et al. 2022. "Genomic Characterization of Metastatic Patterns from Prospective Clinical Sequencing of 25,000 Patients." *Cell* 185 (3): 563–75.e11.
Novakovsky, German, Oriol Fornes, Manu Saraswat, Sara Mostafavi, and Wyeth W. Wasserman. 2022. "ExplaiNN: Interpretable and Transparent Neural Networks for Genomics." *bioRxiv*. https://doi.org/10.1101/2022.05.20.492818.
Olmeda, David, Daniela Cerezo-Wallis, Erica Riveiro-Falkenbach, Paula C. Pennacchi, Marta Contreras-Alcalde, Nuria Ibarz, Metehan Cifdaloz, et al. 2017. "Whole-Body Imaging of Lymphovascular Niches Identifies Pre-Metastatic Roles of Midkine." *Nature* 546 (7660): 676–80.
Ouyang, Zhengqing, Qing Zhou, and Wing Hung Wong. 2009. "ChIP-Seq of Transcription Factors Predicts Absolute and Differential Gene Expression in Embryonic Stem Cells." *Proceedings of the National Academy of Sciences of the United States of America* 106 (51): 21521–26.
Paliou, Christina, Philine Guckelberger, Robert Schöpflin, Verena Heinrich, Andrea Esposito, Andrea M. Chiariello, Simona Bianco, et al. 2019. "Preformed Chromatin Topology Assists Transcriptional Robustness of *Shh* during Limb Development." *Proceedings of the National Academy of Sciences of the United States of America* 116 (25): 12390–99.
Pique-Regi, Roger, Jacob F. Degner, Athma A. Pai, Daniel J. Gaffney, Yoav Gilad, and Jonathan K. Pritchard. 2011. "Accurate Inference of Transcription Factor Binding from DNA Sequence and Chromatin Accessibility Data." *Genome Research* 21 (3): 447–55.
Pliner, Hannah A., Jonathan S. Packer, José L. McFaline-Figueroa, Darren A. Cusanovich, Riza M. Daza, Delasa Aghamirzaie, Sanjay Srivatsan, et al. 2018. "Cicero Predicts Cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data." *Molecular Cell* 71 (5): 858–71.e8.
Pogo, B. G., V. G. Allfrey, and A. E. Mirsky. 1966. "RNA Synthesis and Histone Acetylation during the Course of Gene Activation in Lymphocytes." *Proceedings of the National Academy of Sciences of the United States of America* 55 (4): 805–12.

Polyak, Kornelia, and Robert A. Weinberg. 2009. “Transitions between Epithelial and Mesenchymal States: Acquisition of Malignant and Stem Cell Traits.” *Nature Reviews. Cancer* 9 (4): 265–73.
Quinn, Jeffrey J., Matthew G. Jones, Ross A. Okimoto, Shigeki Nanjo, Michelle M. Chan, Nir Yosef, Trever G. Bivona, and Jonathan S. Weissman. 2021. “Single-Cell Lineages Reveal the Rates, Routes, and Drivers of Metastasis in Cancer Xenografts.” *Science* 371 (6532). https://doi.org/10.1126/science.abc1944.
Rada-Iglesias, Alvaro, Ruchi Bajpai, Tomek Swigut, Samantha A. Brugmann, Ryan A. Flynn, and Joanna Wysocka. 2011. “A Unique Chromatin Signature Uncovers Early Developmental Enhancers in Humans.” *Nature* 470 (7333): 279–83.
Raj, Arjun, and Alexander van Oudenaarden. 2008. “Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences.” *Cell* 135 (2): 216–26.
Ramirez, Ricardo N., Nicole C. El-Ali, Mikayla Anne Mager, Dana Wyman, Ana Conesa, and Ali Mortazavi. 2017. “Dynamic Gene Regulatory Networks of Human Myeloid Differentiation.” *Cell Systems* 4 (4): 416–29.e3.
Raum, Jeffrey C., Kevin Gerrish, Isabella Artner, Eva Henderson, Min Guo, Lori Sussel, Jonathan C. Schisler, Christopher B. Newgard, and Roland Stein. 2006. “FoxA2, Nkx2.2, and PDX-1 Regulate Islet Beta-Cell-Specific mafA Expression through Conserved Sequences Located between Base Pairs -8118 and -7750 Upstream from the Transcription Start Site.” *Molecular and Cellular Biology* 26 (15): 5735–43.
Riolo, Rick, Ekaterina Vladislavleva, Marylyn D. Ritchie, and Jason H. Moore. 2013. *Genetic Programming Theory and Practice X*. Springer Science & Business Media.
Robertson, Gordon, Martin Hirst, Matthew Bainbridge, Misha Bilenky, Yongjun Zhao, Thomas Zeng, Ghia Euskirchen, et al. 2007. “Genome-Wide Profiles of STAT1 DNA Association Using Chromatin Immunoprecipitation and Massively Parallel Sequencing.” *Nature Methods* 4 (8): 651–57.
Robinson, James T., Helga Thorvaldsdóttir, Douglass Turner, and Jill P. Mesirov. 2020. “Igv.js: An Embeddable JavaScript Implementation of the Integrative Genomics Viewer (IGV).” *bioRxiv*. https://doi.org/10.1101/2020.05.03.075499.
Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. 2011. “Integrative Genomics Viewer.” *Nature Biotechnology* 29 (1): 24–26.
Samstein, Robert M., Aaron Arvey, Steven Z. Josefowicz, Xiao Peng, Alex Reynolds, Richard Sandstrom, Shane Neph, et al. 2012. “Foxp3 Exploits a Pre-Existent Enhancer Landscape for Regulatory T Cell Lineage Specification.” *Cell* 151 (1): 153–66.
Sanford, Eric M., Benjamin L. Emert, Allison Coté, and Arjun Raj. 2020. “Gene Regulation Gravitates toward Either Addition or Multiplication When Combining the Effects of Two Signals.” *eLife* 9 (December). https://doi.org/10.7554/eLife.59388.
Satpathy, Ansuman T., Jeffrey M. Granja, Kathryn E. Yost, Yanyan Qi, Francesca Meschi, Geoffrey P. McDermott, Brett N. Olsen, et al. 2019. “Massively Parallel Single-Cell Chromatin Landscapes of Human Immune Cell Development and Intratumoral T Cell Exhaustion.” *Nature Biotechnology* 37 (8): 925–36.

Schep, Alicia N., Beijing Wu, Jason D. Buenrostro, and William J. Greenleaf. 2017. "chromVAR: Inferring Transcription-Factor-Associated Accessibility from Single-Cell Epigenomic Data." *Nature Methods* 14 (10): 975–78.
Schones, Dustin E., Kairong Cui, Suresh Cuddapah, Tae-Young Roh, Artem Barski, Zhibin Wang, Gang Wei, and Keji Zhao. 2008. "Dynamic Regulation of Nucleosome Positioning in the Human Genome." *Cell* 132 (5): 887–98.
Shaffer, Sydney M., Margaret C. Dunagin, Stefan R. Torborg, Eduardo A. Torre, Benjamin Emert, Clemens Krepler, Marilda Beqiri, et al. 2017. "Rare Cell Variability and Drug-Induced Reprogramming as a Mode of Cancer Drug Resistance." *Nature* 546 (7658): 431–35.
Sharma, Sreenath V., Diana Y. Lee, Bihua Li, Margaret P. Quinlan, Fumiyuki Takahashi, Shyamala Maheswaran, Ultan McDermott, et al. 2010. "A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations." *Cell* 141 (1): 69–80.
Slattery, Matthew, Todd Riley, Peng Liu, Namiko Abe, Pilar Gomez-Alcala, Iris Dror, Tianyin Zhou, et al. 2011. "Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox Proteins." *Cell* 147 (6): 1270–82.
Smith, Richard, Leah A. Owen, Deborah J. Trem, Jenny S. Wong, Jennifer S. Whangbo, Todd R. Golub, and Stephen L. Lessnick. 2006. "Expression Profiling of EWS/FLI Identifies NKX2.2 as a Critical Target Gene in Ewing's Sarcoma." *Cancer Cell* 9 (5): 405–16.
Song, Xue, Zhongyun Zhao, Beth Barber, Amanda M. Farr, Boris Ivanov, and Marilyn Novich. 2015. "Overall Survival in Patients with Metastatic Melanoma." *Current Medical Research and Opinion* 31 (5): 987–91.
Spitz, François, and Eileen E. M. Furlong. 2012. "Transcription Factors: From Enhancer Binding to Developmental Control." *Nature Reviews. Genetics* 13 (9): 613–26.
Starks, Rebekah R., Anilisa Biswas, Ashish Jain, and Geetu Tuteja. 2019. "Combined Analysis of Dissimilar Promoter Accessibility and Gene Expression Profiles Identifies Tissue-Specific Genes and Actively Repressed Networks." *Epigenetics & Chromatin* 12 (1): 16.
Strahl, B. D., and C. D. Allis. 2000. "The Language of Covalent Histone Modifications." *Nature* 403 (6765): 41–45.
Strogatz, Steven, Mark Friedman, A. John Mallinckrodt, and Susan McKay. 1994. "Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering." *Computer Physics Communications* 8 (5): 532.
Sung, Myong-Hee, Songjoon Baek, and Gordon L. Hager. 2016. "Genome-Wide Footprinting: Ready for Prime Time?" *Nature Methods* 13 (3): 222–28.
Symmons, Orsolya, and Arjun Raj. 2016. "What's Luck Got to Do with It: Single Cells, Multiple Fates, and Biological Nondeterminism." *Molecular Cell* 62 (5): 788–802.
Taniguchi, Yasushi. 2016. "The Bromodomain and Extra-Terminal Domain (BET) Family: Functional Anatomy of BET Paralogous Proteins." *International Journal of Molecular Sciences* 17 (11). https://doi.org/10.3390/ijms17111849.
Tena, Juan J., and José M. Santos-Pereira. 2021. "Topologically Associating Domains and Regulatory Landscapes in Development, Evolution and Disease." *Frontiers in Cell and Developmental Biology* 9 (July): 702787.

Teves, Sheila S., Luye An, Anders S. Hansen, Liangqi Xie, Xavier Darzacq, and Robert Tjian. 2016. “A Dynamic Mode of Mitotic Bookmarking by Transcription Factors.” *eLife* 5 (November). https://doi.org/10.7554/eLife.22280.
Thurman, Robert E., Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T. Maurano, Eric Haugen, Nathan C. Sheffield, et al. 2012. “The Accessible Chromatin Landscape of the Human Genome.” *Nature* 489 (7414): 75–82.
Torre, Eduardo A., Eri Arai, Sareh Bayatpour, Connie L. Jiang, Lauren E. Beck, Benjamin L. Emert, Sydney M. Shaffer, et al. 2021. “Genetic Screening for Single-Cell Variability Modulators Driving Therapy Resistance.” *Nature Genetics* 53 (1): 76–85.
Torre-Ubieta, Luis de la, Jason L. Stein, Hyejung Won, Carli K. Opland, Dan Liang, Daning Lu, and Daniel H. Geschwind. 2018. “The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis.” *Cell* 172 (1-2): 289–304.e18.
Turner, Noel, Olivia Ware, and Marcus Bosenberg. 2018. “Genetics of Metastasis: Melanoma and Other Cancers.” *Clinical & Experimental Metastasis* 35 (5-6): 379–91.
Vaishnav, Eeshit Dhaval, Carl G. de Boer, Jennifer Molinet, Moran Yassour, Lin Fan, Xian Adiconis, Dawn A. Thompson, Joshua Z. Levin, Francisco A. Cubillos, and Aviv Regev. 2022. “The Evolution, Evolvability and Engineering of Gene Regulatory DNA.” *Nature* 603 (7901): 455–63.
Voss, Ty C., R. Louis Schiltz, Myong-Hee Sung, Paul M. Yen, John A. Stamatoyannopoulos, Simon C. Biddie, Thomas A. Johnson, Tina B. Miranda, Sam John, and Gordon L. Hager. 2011. “Dynamic Exchange at Regulatory Elements during Chromatin Remodeling Underlies Assisted Loading Mechanism.” *Cell* 146 (4): 544–54.
Weintraub, H., and M. Groudine. 1976. “Chromosomal Subunits in Active Genes Have an Altered Conformation.” *Science* 193 (4256): 848–56.
West, Jason A., April Cook, Burak H. Alver, Matthias Stadtfeld, Aimee M. Deaton, Konrad Hochedlinger, Peter J. Park, Michael Y. Tolstorukov, and Robert E. Kingston. 2014. “Nucleosomal Occupancy Changes Locally over Key Regulatory Regions during Cell Differentiation and Reprogramming.” *Nature Communications* 5 (August): 4719.
Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the Tidyverse.” *Journal of Open Source Software* 4 (43): 1686.
Williamson, Iain, Lauren Kane, Paul S. Devenney, Ilya M. Flyamer, Eve Anderson, Fiona Kilanowski, Robert E. Hill, Wendy A. Bickmore, and Laura A. Lettice. 2019. “Developmentally Regulated Shh Expression Is Robust to TAD Perturbations.” *Development* 146 (19). https://doi.org/10.1242/dev.179523.
Wu, C., Y. C. Wong, and S. C. Elgin. 1979. “The Chromatin Structure of Specific Genes: II. Disruption of Chromatin Structure during Gene Activity.” *Cell* 16 (4): 807–14.
Wu, Tianzhi, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, et al. 2021. “clusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data.” *Innovation (New York, N.Y.)* 2 (3): 100141.
Xu, X., Z. Yin, J. B. Hudson, E. L. Ferguson, and M. Frasch. 1998. “Smad Proteins Act in Combination with Synergistic and Antagonistic Regulators to Target Dpp Responses to the Drosophila Mesoderm.” *Genes & Development* 12 (15): 2354–70.

Yang, Yajie, Justin Fear, Jianhong Hu, Irina Haecker, Lei Zhou, Rolf Renne, David Bloom, and Lauren M. McIntyre. 2014. “Leveraging Biological Replicates to Improve Analysis in ChIP-Seq Experiments.” *Computational and Structural Biotechnology Journal* 9 (January): e201401002.

Yoshida, Akihiko, Shigeki Sekine, Koji Tsuta, Masashi Fukayama, Koh Furuta, and Hitoshi Tsuda. 2012. “NKX2.2 Is a Useful Immunohistochemical Marker for Ewing Sarcoma.” *The American Journal of Surgical Pathology* 36 (7): 993–99.

Yu, Guangchuang, Li-Gen Wang, and Qing-Yu He. 2015. “ChIPseeker: An R/Bioconductor Package for ChIP Peak Annotation, Comparison and Visualization.” *Bioinformatics* 31 (14): 2382–83.

Zaidi, Sayyed K., Daniel W. Young, Martin A. Montecino, Jane B. Lian, Andre J. van Wijnen, Janet L. Stein, and Gary S. Stein. 2010. “Mitotic Bookmarking of Genes: A Novel Dimension to Epigenetic Control.” *Nature Reviews. Genetics* 11 (8): 583–89.

Zaret, Kenneth S. 2020. “Pioneer Transcription Factors Initiating Gene Network Changes.” *Annual Review of Genetics* 54 (November): 367–85.

Zhang, Yanjun, Zhongxing Sun, Junqi Jia, Tianjiao Du, Nachuan Zhang, Yin Tang, Yuan Fang, and Dong Fang. 2021. “Overview of Histone Modification.” In *Histone Mutations and Cancer*, edited by Dong Fang and Junhong Han, 1–16. Singapore: Springer Singapore.

Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, et al. 2008. “Model-Based Analysis of ChIP-Seq (MACS).” *Genome Biology* 9 (9): R137.

Zhou, Yeqiao, Jelena Petrovic, Jingru Zhao, Wu Zhang, Ashkan Bigdeli, Zhen Zhang, Shelley L. Berger, Warren S. Pear, and Robert B. Faryabi. 2022. “EBF1 Nuclear Repositioning Instructs Chromatin Refolding to Promote Therapy Resistance in T Leukemic Cells.” *Molecular Cell* 82 (5): 1003–20.e15.