1-13-2024

A Test of the Pioneer Factor Hypothesis for Silent Gene Activation

Jeffrey Hansen
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Computational and Systems Biology

Dissertation Examination Committee:
Barak Cohen, Chair
Robi Mitra
Samantha Morris
Gary Stormo
Ting Wang

A Test of the Pioneer Factor Hypothesis for Silent Gene Activation

by
Jeffrey L. Hansen

A dissertation presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2024
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgements

Thank you, Dr. Barak Cohen for lessons at the whiteboard, constructive deliberation at lab meetings, and casual scientific artistry in Arkansas, Wisconsin, and the Scottish Arms. Your science is clear, your mentorship is warm, and your friendship is dear.

Thank you, Cohen Lab, and especially Clarice and Siqi. You helped me mope over donuts, but then took me back into lab to help troubleshoot my experimental woes.

Thank you, Drs. Ting Wang, Samantha Morris, Gary Stormo, and Rob Mitra for serving on my thesis committee. I am so lucky to have collaborated with such a brilliant and fun group.

Thank you, Dr. Wayne Yokoyama and the entire MSTP for the opportunity to study at WashU and for helping me to navigate the complexity of the MD-PhD program.

Thank you, Jess Hoisington-Lopez and ML Crosby for sequencing so many of my libraries.

Thank you, Bonnie Dee, Pat Winkler, and Shirley McKinney for your joyful support in ordering supplies, paying for conference travel, and locating lab equipment.

Thank you, friends, and family for helping me to remember the sun despite the occasional clouds. There was so much sunshine!

And **thank you** to my funding sources (T32GM007200, T32HG000045, R01GM92910).

Love,

Jeff

Washington University in St. Louis

May 2024

ABSTRACT OF THE DISSERTATION

A Test of the Pioneer Factor Hypothesis for Silent Gene Activation

by

Jeffrey L. Hansen

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2024

Professor Barak A. Cohen, Chair

Transcription factors (TFs) activate silent genes by binding to and opening heterochromatic instances of their motifs. While we rely on this process for cellular reprogramming, we have an incomplete understanding of which TFs are capable of recognizing inaccessible instance of their motifs and what parameters are important for this activity. My thesis work aimed to address these two questions.

The leading model for silent gene activation is the pioneer factor hypothesis (PFH). The PFH states that pioneer factors (PFs) are qualitatively unique TFs that can bind to and open DNA and subsequently recruit non-pioneer factors (nonPFs) to activate expression. We tested the predictions of the PFH by ectopically expressing a canonical PF FOXA1 and nonPF HNF4A in K562 blood cells. While we expected that only FOXA1 would bind inaccessible motifs and that neither TF would activate tissue-specific gene expression, we found that both TFs independently bound, opened, and activated tissue-specific loci. When we examined what may control such "pioneer activity," we found that motif content, TF concentration, and TF binding strength were all important factors.

Having shown that pioneer activity may not be a qualitative trait restricted to just a few TFs, we sought to develop a quantitative metric. Because pioneer activity is essentially "TF binding at hard-to-bind sites," we suggest that a measure of pioneer activity should capture the relative difference in a TF's ability to bind at accessible versus inaccessible DNA. We estimated a parameter related to a TF's K_d by using doxycycline (dox) induction as a proxy for TF concentration. We call this term the TF's dox_{50}. We propose that the average difference of a TF's dox_{50} between accessible and inaccessible binding sites is a measure of its pioneer activity. We call this term the TF's Δdox_{50}. The lower a TF's Δdox_{50}, the stronger its pioneer activity. To demonstrate the feasibility of this metric, we induced FOXA1 and HNF4A across a 1,000-fold range, measured binding, fit binding curves at tens of thousands of loci, and then extracted dox_{50}s. We show that HNF4A has a smaller Δdox_{50} than FOXA1, which suggests it has stronger pioneer activity. We also show that FOXA1 has a smaller Δdox_{50} at sites that have more copies of its motif, which suggests that strong motif content can boost pioneer activity.

Altogether we propose that every TF likely has some degree of pioneer activity that depends on its affinity for any given location, the concentration at which it is expressed, and the motif content at each target site. We hope that future work will characterize more TFs' pioneer activity, making Δdox_{50} a useful quantitative metric to describe pioneer activity.

Chapter 1 – Introduction

Every cell in the human body contains the same set of genetic instructions. Yet our brains think, and our hearts beat. Somehow our cells selectively read or ignore portions of the 3.2 billion base pair code to achieve specialization. We took a major step in understanding how when we assembled the first full sequence of the human genome in 2001 (Venter et al. 2001) but soon after realized that less than 2% of the sequence codes for the genes that make the proteins that make our cells (Elkon and Agami 2017). There are many outstanding questions related to how the rest of the "regulatory DNA" somehow ensures that each cell type has the appropriate genes turned on at the appropriate times. Some of these questions include: how is DNA arranged three-dimensionally within the nucleus? How is specificity achieved from a repeating sequence of just four nucleotides? Or how might we predict disease occurrence or severity by examining the DNA sequence of key genes? My lab mates are actively performing experiments to answer these questions and have already begun to uncover important principles. Alongside them, I have designed my work to combine my interests in gene regulation and human disease, especially in the spirit of my dual MD-PhD training.

One especially interesting application is the question of how to turn on genes that are not currently being used. We know that this occurs naturally through development as cells turn on new genes so that they can begin to differentiate into various specializations. But we also have discovered that we can turn on lineage genes in the wrong cell type through a process called cellular reprogramming. I became interested in reprogramming by way of my long interest in novel treatments or cures for type 1 diabetes. Early in my graduate training, I read about new

work that could convert pancreatic alpha cells (those that use glucagon to increase blood glucose levels) into pancreatic beta cells (those that use insulin to decrease blood glucose levels). I will later discuss the details for how this conversion works, but it relies on important proteins in the cell called transcription factors (TFs). TFs are normally responsible for tuning cell type-specific gene expression but it turns out that they are also capable of activating other lineage's genes and inducing cellular conversions. I refer to this process as silent gene activation. The process is incompletely understood, leads to inefficient cellular conversions, and thus presents an exciting opportunity for meaningful progress. I hope that my contributions may allow us not only to better understand the mechanisms of gene regulation but also to ultimately develop even better therapies for type 1 diabetes in the near future.

1.1 – Cell-type specific gene expression

Only a small fraction of the 3.2 billion base pairs in the genome code for the ~20,000 genes that make the proteins that build our cells. It is estimated that 3,000-8,000 of these genes perform common jobs such as growth, division, or energy consumption that are important to all cell types (J. Zhu et al. 2008; Ramsköld et al. 2009), but this only leaves 15,000 or so to create the cell-type specialization that makes humans so complex. If there are approximately 200 cell types in the human body (Heintzman et al. 2009), this means that there are approximately 75 genes that are specific to each cell type. This simple calculation is supported by studies that have used gene expression across cell types to generate lists of cell type-specific genes. One such resource, the Human Protein Atlas, lists for example that there are 242 liver-specific genes and 120 intestine-specific genes (Uhlén et al. 2015). The liver genes perform jobs related to energy metabolism or

blood detoxification and the intestine genes perform jobs related to digestion and food absorption. In cell types that do not perform these jobs, these genes are unnecessary and are stored away.

The mechanism by which these genes are stored was first theorized by a German botanist named Emil Heitz when he observed compacted chromosomes, hypothesized that they may be related to genetic silencing, and coined the terms “heterochromatin” and “euchromatin” to describe the silent and active states, respectively (Heitz 1928). His hypothesis was soon verified when it was found that heterochromatin can silence nearby genes (Schultz 1936). Later work revealed that the building block of heterochromatin is approximately 150 base pairs of DNA wrapped around an octamer of histone proteins called a “nucleosome;” this structure has casually been referred to as “beads on a string” (Woodcock, Safer, and Stanchfield 1976; Kornberg 1974; Oudet, Gross-Bellard, and Chambon 1975). Nucleosomes sequester DNA to sterically hinder access by other proteins. The structure can be compacted even further, making it even more silent, by packing nucleosomes into repeating arrays sometimes referred to as the “30 nanometer fiber.”

Since these discoveries, we now know that heterochromatin is the result of competition between silencing and activation processes that upon establishment locks away unnecessary genes into a silent state (Elgin 1996). Importantly, it is not only the silent genes that are included within heterochromatin but also the proximal regulatory DNA. This regulatory DNA is the substrate to which TFs bind and then recruit transcriptional machinery to activate nearby genes. Thus heterochromatin effectively shields TFs from reaching their targets within regulatory DNA. Further studies have sub-classified heterochromatin into facultative and constitutive (Mayran et

al. 2018) and have identified myriad covalent modifications made to the histone proteins that may have functional consequences and are often used to predict activity of nearby genes (Ernst and Kellis 2012). New technologies have also been invented to study heterochromatin. These technologies treat nuclei with enzymes such as DNase (L. Song and Crawford 2010), MNase (Schones et al. 2008), or transposase (Buenrostro et al. 2015) and then use high-throughput sequencing to identify which parts of the genome were "accessible" or "inaccessible" to the enzymes. Thus the terms silent, inaccessible, nucleosomal, and heterochromatic are all used to describe DNA that has been packaged and stored away.

In healthy cells, the steric hindrance that heterochromatin provides is robust and the division between euchromatin and heterochromatin is stable; liver cells continue to detoxify and intestine cells continue to absorb. Patterns of chromatin accessibility are so cell-type specific that they have recently been demonstrated to define cell types when measured at single cell resolution (Lareau et al. 2019). Despite the seeming permanence and stability of heterochromatic gene silencing, we know there are cases when TFs activate silent genes. In natural development, lineage-determining TFs must find their lineage-specific genes within heterochromatin and turn them on. And in disease, malignant states in cancer can disrupt heterochromatin and turn on previously silenced genes (Robson et al. 1981). Heterochromatin may not be as permanent or inaccessible as perhaps once thought. This realization and others have led to new strategies to specifically disrupt heterochromatin to turn on genes of different lineages.

1.2 – Reprogramming one cell type into another

Gurdon et al. first demonstrated the potential of manipulating the plasticity of the genome when they successfully transplanted somatic gut frog nuclei into enucleated ova to grow healthy, adult frogs (Gurdon 1962). This work shows that even a fully differentiated gut cell has the potential to essentially go "back-in-time" to a pluripotent state (dedifferentiate) and then progress forward through new developmental trajectories (redifferentiate) into each necessary cell type of the adult frog. Researchers demonstrated that the same reprogramming strategy is feasible in mammals when they grew viable lamb offspring (Dolly) from differentiated adult tissue (Wilmut et al. 1997). While these innovations were incredibly novel in the research setting, the process of dedifferentiation and redifferentiation occurs naturally in zebrafish, who are able to regenerate resected portions of their hearts (Jopling et al. 2010).

26 years after Gurdon et al. showed that we could reprogram cells, Tapscott et al. showed that the expression of a single muscle-specific TF MYOD1 was sufficient to reprogram fibroblasts into myoblasts (Tapscott et al. 1988; Davis, Weintraub, and Lassar 1987). While TFs were traditionally known to be responsible for gene regulation in their native tissue, these experiments revealed a novel ability for TFs to somehow overcome heterochromatin's steric hindrance and reactivate silent genes of other lineages. Since then, work to understand how TFs control dedifferentiation, redifferentiation, and transdifferentiation (reprogramming from one differentiated cell type to another without using a stem cell-like intermediary) has rapidly expanded. We now have cocktails of TFs that can convert adult fibroblasts backwards in time to

induced pluripotent stem cells (Takahashi and Yamanaka 2006) as well as cocktails to convert fibroblasts into neuronal, hepatic, cardiac, and other lineages (Samantha A. Morris 2016).

The innovation of these works and the potential for using reprogrammed cells for research and medical purposes earned John Gurdon (reprogrammed the frog) and Shinya Yamanaka (created induced pluripotent stem cells) the Nobel Prize in medicine in 2012. For use in the lab, reprogrammed cells offer researchers the ability to directly create cells that are often challenging or time-consuming to harvest. One good example is the ability to reprogram aged fibroblasts into aged neurons, allowing for the study of age-related neurodegenerative disease (Huh et al. 2016). And for use in the clinic, reprogrammed cells could replenish those lost to disease or damage; this is especially useful when healthy, actively dividing cells such as microglia or cardiac fibroblasts reside in proximity to diseased, post-mitotic cells such as neurons or cardiomyocytes. Following a heart attack, we could convert cardiac fibroblasts into cardiomyocytes (Chang et al. 2019; Qian et al. 2012; K. Song et al. 2012; Jayawardena et al. 2015; Ieda et al. 2010). Following a stroke, we could convert microglia (Matsuda et al. 2018) or astrocytes (Su et al. 2014) into functional neurons. And when the autoimmune response driving type 1 diabetes depletes pancreatic beta cells, we can reprogram nearby alpha cells, which serendipitously exhibit hypo-immunogenicity that allows some evasion of the autoimmune attack (Furuyama et al. 2019; Thorel et al. 2010; Pagliuca et al. 2014; Velazco-Cruz et al. 2018).

1.3 – Inefficiencies of reprogramming and challenges of binding at heterochromatic sites

While it is remarkable that we can induce these conversions, there are two major impediments to realizing the full potential. First, few cells in the initial starting population make it to the desired end point. The conversion percentage depends on the reprogramming cocktail but hovers around 10% across different tissue types (Ieda et al. 2010; Zhao et al. 2015; Vierbuchen et al. 2010). In fact the Nobel Prize winning Yamanaka TFs converted less than 0.1% of the initial fibroblasts to induced embryonic stem cells (Takahashi and Yamanaka 2006). And second, the cells that do make it through the conversion process are often either hybrid cells that express gene signatures common to both the starting and desired cell type (part fibroblast, part cardiomyocyte/neuron) (Ieda et al. 2010; Manandhar et al. 2017) or become stuck in a state that is developmentally immature to the desired cell type (Biddy et al. 2018).

These shortcomings arise because the TFs in reprogramming cocktails incompletely activate the necessary genes. I speculate that this is not a reflection on the general ability of TFs to activate silent genes but rather a reflection of our inadequate understanding of which TFs to employ, how to employ them, and when to employ them. Instead of first understanding the mechanism by which TFs activate silent genes and then using this information to rationally design reprogramming cocktails from the bottom up, most cocktails are designed in large screening-based techniques. Some strategies test many TFs in parallel for their ability to turn on lineage specific genes (Ng et al. 2021). Others start with a large set of TFs known to be important in a certain lineage and then drop one TF out at a time to find the smallest sufficient set (Sekiya and

Suzuki 2011). Then, upon selection of the cocktail, the researchers will draw conclusions from genome-wide binding, accessibility, or other epigenetic data in an attempt to explain the TFs behavior (Horisawa et al. 2020). From these data, TFs are labeled as more or less important for the reprogramming cocktail. But rarely are the conclusions about these TFs directly tested in a one-by-one process and so our basic understanding of the mechanism underpinning cellular reprogramming remains incompletely understood.

At this point I recognized multiple opportunities for how I could fit my work into these challenges. First, I could combine my interests in the human genome and cellular reprogramming. Second, I could leverage our lab's expertise in gene regulation to further the field's understanding of silent gene activation. And third, I could perhaps provide an example of a simple, effective way to test TFs for their ability to activate genes. The big question of my thesis work thus became:

How do transcription factors activate silent genes?

Alongside the problem of activating the necessary genes of the desired final cell type, there may also be a need to further understand how to shut off the transcriptional program of the initial cell population. Not only do we see that some genes from the target cell type are never activated, but we also see that some genes from the initial cell type are never silenced (Manandhar et al. 2017). And data suggests that strategies to ensure that these genes are silenced may improve cellular reprogramming (Zhao et al. 2015). That said, most of the work to date that focuses on inhibition of gene expression during reprogramming utilizes small molecules or microRNAs (Zhao et al.

2015; Yoo et al. 2011; Muraoka et al. 2014), not TFs. Less is known about the repressive capabilities of mammalian TFs. I also speculate that like the normal process of development, sufficient activation of one lineage's genes may create enough positive feedback that inhibition may not be necessary. For these reasons I have limited the scope of my work solely to the question of silent gene activation.

The challenge that TFs need to overcome in order to activate silent genes is that the short sequences to which they specifically bind (their "motifs") are wrapped around histones and compacted into heterochromatin. As mentioned above, the mechanism by which silent genes are kept silent is by heterochromatic compaction of the sites to which TFs bind. Without the ability to bind, the TFs are unable to activate nearby genes. Multiple different methodologies have shown that TFs bind more weakly, or not at all, to inaccessible instances of their motifs. Electrophoresis mobility shift assays (EMSAs) in the presence or absence of nucleosomes showed that for nearly all of the TFs tested, the TFs had a larger K_d (weaker binding) when binding to nucleosomal DNA (Garcia et al. 2019). And in vitro nucleosome reconstitution assays where DNA is artificially compacted into heterochromatin showed that heterochromatin blocks the binding of some TFs (Lisa Ann Cirillo et al. 2002).

The heterochromatin did not inhibit binding of all of the tested TFs, though. The authors found that some TFs, such as the liver TF FOXA1, could bind and decompact heterochromatin and fatefully coined the term "pioneer factors" to refer to the TFs that could "pioneer" unexplored regions of the genome (Lisa Ann Cirillo et al. 2002). Later, the TFs that had been classified as unable to bind heterochromatic motifs were called "settler TFs" in order to complete the perhaps

ill-fated metaphor (Sherwood et al. 2014). While this qualitative distinction may have been too simplistic given previous work that showed TFs may bind nucleosomal DNA in a non-specific, transient process (Polach and Widom 1995, 1996), the pioneer factor hypothesis (PFH) quickly became the leading model for how TFs activate silent genes (Lisa Ann Cirillo et al. 2002; Iwafuchi-Doi and Zaret 2014).

1.4 – The Pioneer Factor Hypothesis of silent gene activation

The PFH has two components. First, it establishes that there are qualitatively different types of TFs. Pioneer factors (PFs) can bind to their motifs within heterochromatin and non-pioneer factors (nonPFs) cannot. Cirillo et al. demonstrated this by showing that only some TFs (the PFs) could bind in vitro to a compacted sequence of regulatory DNA specific to the liver gene *ALB* (Lisa Ann Cirillo et al. 2002). And second, gene activation requires a sequential process where PFs bind first, create new local accessibility, and then recruit nonPFs to activate gene expression. This claim is supported by some data that show that genome-wide patterns of nonPFs are affected by PF binding (Horisawa et al. 2020) and other data that show that TF binding clusters at sites where PFs bind (Iwafuchi-Doi and Zaret 2014).

The PFH's second component likely relates to TFs' involvement within the cascades of positive feedback during development. TFs activate batteries of genes necessary for development and combinations of TFs are thought of as one way to achieve tissue specificity. The earlier expressed TFs must then be the ones to initiate the cascade. The two most well-studied PFs are two such developmentally important TFs. FOXA1 is critical for liver bud development (Lee et

al. 2005). And Zelda is necessary both for *Drosophila* zygotic genome activation (McDaniel et al. 2019) and for important transitions during the development of neuroblasts (Larson et al. 2021). It is hard to know though whether the order of TF expression was set up because TFs such as FOXA1 and Zelda are the only ones capable of binding at heterochromatin, or for other non-binding related reasons. I suspect that simply being expressed early in development may have contributed to some TFs being labeled as PFs.

Since the first PF paper, others have designed similar experiments to classify more TFs. These experiments can generally be categorized as in vivo ectopic expression systems, in vivo synthetic screens, or in vitro synthetic screens. The in vivo ectopic expression systems express a TF, measure its genome-wide binding patterns, correlate the patterns with accessibility and chromatin modifications, and then classify the tested TF as a PF or nonPF (Donaghey et al. 2018; Wapinski et al. 2013). The in vivo synthetic screens integrate large numbers of synthetic regulatory elements into the genome, compact the elements within heterochromatin, and then observe which elements induce decompaction (Yan, Chen, and Bai 2018; Hammelman et al. 2020; Sherwood et al. 2014). And the in vitro synthetic screens test large numbers of synthetic regulatory elements on artificial nucleosome arrays outside of the cellular environment (Yu and Buck 2019).

Once there was a list of PFs and nonPFs, there became interest in understanding what endowed PFs the ability to bind heterochromatic instances of their motifs. The poster child is FOXA1. FOXA1 is a liver PF, was the first TF to be named a PF in the above study (Lisa Ann Cirillo et al. 2002), and has a winged-helix DNA-binding domain that is similar to the three-dimensional

conformation of the linker histone protein H1 (Clark et al. 1993; Ramakrishnan et al. 1993). H1 is one component of the nucleosome complex and is partially responsible for the further compaction of nucleosomes into the 30nm fiber. It was proposed that FOXA1 could outcompete nucleosomes for the underlying DNA better than other TFs because of its similar shape. Following this work, a larger scale study searched crystal structures of PFs and nonPFs for structural differences and claimed that the PF binding ability resides in a short alpha helix within the TF's DNA-binding domain (Garcia et al. 2019). While some PFs in this study did contain this structural motif, some PFs did not and some nonPFs did; the motif was neither necessary nor sufficient.

We have yet to find a single unifying structural motif to classify PFs from nonPFs. Instead each PF uses a different strategy to bind heterochromatic DNA. Each strategy relates to a unique type binding behavior between TF and nucleosomal DNA. First, some TFs such as P53 require that their motifs reside at the position where DNA enters or exits the nucleosome core; at these positions, DNA is more accessible to TF intrusion (Yu and Buck 2019). Second, TFs may collaborate (without physical interaction) to pull DNA away from the nucleosome (Mirny 2010). The optimal spacing appears to be ~74 base pairs so that one TF is positioned at the entry point into the nucleosome and the other TF is positioned on the same side of the nucleosome (Moyle-Heyrman, Tims, and Widom 2011). Third, some TFs have DNA-binding domains that allow them to bind to partial motifs so that binding is still possible when part of the DNA strand is occluded by the nucleosome (Soufi et al. 2015). And fourth, some TFs rely on arrays of motifs that are found 10-15 base pairs apart from one another; this periodicity corresponds to consecutive major grooves of the DNA (Casey et al. 2018). A nice study aimed to measure these

binding modalities in parallel, queried 220 TFs, and identified five types: spanning the 2 nucleosomal gyres, binding at the dyad, binding at the end, binding periodically, or having an orientation preference (F. Zhu et al. 2018).

The breadth of strategies that different TFs use to bind nucleosomal DNA and the finding that motif positioning can either allow or prohibit pioneer activity seems to suggest that pioneer activity may be a characteristic of a certain state rather than certain TFs. In the same high-throughput binding study mentioned above, the authors found that most of their 220 tested TFs had some ability to bind nucleosomal DNA (F. Zhu et al. 2018). Prior to the coining of the "pioneer factor" term, I believe that we were headed towards an alternative model in which competition between non-specific TF interactions and histones determined the DNA's accessibility. If TFs win, the DNA is accessible and the nearby genes are active. If the histones win, the DNA is inaccessible and the nearby genes are silent. This model was termed "collaborative competition" and the authors argue that it can quantitatively explain nucleosomal binding (Miller and Widom 2003; Polach and Widom 1996). This model versus the PFH sets up a nice distinction between whether pioneer activity is a qualitative trait limited to a select few TFs or rather a quantitative one that any TF can exhibit given the right conditions.

It may also be that TFs are unnecessary to decompact nucleosomal DNA. DNA may become transiently free from histones all on its own. One good example is that DNA is temporarily naked of nucleosomes at its replication fork. Cell replication has been shown to be required for some TFs to access their heterochromatic motifs (Ramachandran and Henikoff 2016; Yan, Chen, and Bai 2018), suggesting that TFs are probably capitalizing upon the transiently exposed DNA.

DNA is also known to transiently "breathe" away from nucleosomes (Polach and Widom 1995); this could be part of how P53 binds nucleosomes, as previously described (Yu and Buck 2019). And finally we know that there exist pervasive multivalent chromatin remodeling complexes that regularly rearrange nucleosomes and chromatin modifications that likely have effects on how TFs can access their motifs (Stephanie A. Morris et al. 2014).

1.5 – Problems with the Pioneer Factor Hypothesis

Assuming that the transient mechanisms mentioned above are not sufficient and that TFs alone are sometimes necessary to decompact chromatin, the simple binary classification that the PFH makes seems insufficient. My impression is that the PFH has maintained support because data has been selectively highlighted or ignored in order to preserve the neat subdivision of TFs into those that can and those that can't bind nucleosomal DNA. While it is enticing to draw lines between sets of TFs, the oversimplification misses details, clouds our understanding, and will ultimately limit how we select TFs when we seek to ectopically activate silent genes.

Often in the experimental design, other unmentioned factors may account for the behavior of the so-called PFs. The initial authors that coined the term compacted a piece of DNA encoding the *ALB* enhancer within an artificial nucleosome array (Lisa Ann Cirillo et al. 2002). They called the TFs that could bind to the heterochromatic DNA PFs (Lisa Ann Cirillo et al. 2002). The problem with this experimental design is that they did not mention controlling for the TF concentration nor the motif content of the enhancer. Both of these factors will impact the overall occupancy of the TFs at the enhancer. If the concentration of the PF and nonPF or the motif

content at each TF's binding site were not roughly equal, then it is unfair to compare the outcome of the binding experiment.

In vivo experiments are also fraught with covariates. While in vitro experiments have their flaws as mentioned above, they do benefit from their simple design with few moving parts. In contrast, it is often possible to find every potential outcome in the data of an in vivo experiment because of the size of the genome and myriad co-factors and chromatin marks present. The classic in vivo PF experiment ectopically expresses multiple TFs, measures genome-wide binding, and then draws a distinction based on the degree to which each TF can bind inaccessible sites (Donaghey et al. 2018; Wapinski et al. 2013). But in these experiments there also exist at least some inaccessible sites where the nonPF can bind, even if these sites are not the majority. If these inaccessible binding sites happen to reside nearby genes that are important to the TF's lineage, and these genes are activated, then it seems shortsighted to have labeled that TF a nonPF.

Additionally, the data rarely conform to the strict definition of the PFH. The PFH states that PFs can bind to inaccessible motifs and so it would follow that PFs should bind to all of their motifs. While this is perhaps an unrealistic expectation, PFs bind to a small minority of their available motifs (Boller et al. 2016; Lupien et al. 2008), sometimes because of impermissive heterochromatin (Mayran et al. 2018) and other times because binding seems mostly limited to the sites at which a TF binds in its native cell type (Donaghey et al. 2018). There are also cases where PFs behave like nonPFs, and vice versa. Two interesting cases stand out. In the first, the canonical PF FOXA1 requires help from the supposed nonPF estrogen receptor or glucocorticoid receptor in order to bind (Swinstead et al. 2016; Zaret and Carroll 2011). And in the second,

LEXA, a bacterial protein that has no need to evolve the ability to bind nucleosomes, can create new accessibility at heterochromatic sites (Miller and Widom 2003).

Finally, it doesn't make intuitive sense that there would be a requirement for the sequential process of binding, opening, and then recruitment of another factor in order to achieve gene activation. We have known for a long time that TFs can both bind DNA and also recruit the transcriptional machinery (i.e. RNA polymerase) necessary to activate expression. Therefore, if we observe TF binding, then we would expect that the TF could also activate expression of nearby genes. In fact some of the earliest reprogramming experiments relied on the activity of a single TF (MYOD1) to convert fibroblasts into myoblasts (Davis, Weintraub, and Lassar 1987; Choi et al. 1990). It could be possible that the opening (not activation) step requires multiple TFs to outcompete nucleosomes as suggested above (Miller and Widom 2003), but even in this case, it doesn't seem necessary to have multiple unique TFs. Either a TF binding as a homodimer or multiple copies of the same TF binding to an array of spaced motifs could achieve the protein density at an enhancer that may be necessary to outcompete histones.

1.6 – Scope of this dissertation

After reading all of the literature cited above, my hypothesis was that there is nothing qualitatively different about PFs and nonPFs. I predicted that given the right circumstances, I could turn a PF into a nonPF, and vice versa. I initially proposed to test this prediction by integrating a library of regulatory elements into a heterochromatic region of the genome, waiting until those elements became inaccessible, and then expressing TFs whose motifs were in those

elements to see which motifs regained accessibility. My prediction was that low expression of a PF in combination with few or inappropriately spaced target motifs would cause the PF to act like a nonPF. Similarly, a highly expressed nonPF that targeted a sequence with high motif content would act like a PF. The problem that I encountered was that the gene expression readout of genomically silenced elements (i.e. no expression) looks identical to that of elements that were never actually integrated. While I thought that I had integrated elements that were subsequently silenced, I was in fact observing the absence of integrated elements. At that point, I had spent a couple of years trying to establish this method to no avail and decided that it was time to pivot. Coincidentally, around the same time a similar method was published that appeared to encounter similar issues with integrating sequences to be silenced (Hammelman et al. 2020).

After these challenges, I conceded to use a simpler experimental design that would still allow me to study silent gene activation. I realized that the PFH makes clear and testable predictions about ectopically expressed TFs. I also realized that one of my committee members regularly uses PF FOXA1 and nonPF HNF4A to reprogram fibroblasts towards the liver lineage (Biddy et al. 2018). By using FOXA1, I could test the most well studied PF (Lisa Ann Cirillo et al. 2002; Donaghey et al. 2018). By using HNF4A, I could test how a nonPF behaves individually as compared to when expressed with FOXA1 (Horisawa et al. 2020). And by using them together, I could compare my data to that generated from reprogramming experiments that employ the same TFs (Biddy et al. 2018). Finally, and perhaps best of all, none of the involved experiments–lentiviral transduction, RNA-sequencing, ATAC-sequencing, or CUT&Tag–are notoriously challenging to conduct. I have detailed this work in the following two chapters.

I start in Chapter 2 with work that I designed to explicitly test the major predictions that the PFH makes about ectopically expressed PFs and nonPFs. As I mentioned above, the PFH states that: 1) PFs can bind and open inaccessible instances of their motifs while nonPFs cannot, and 2) gene activation requires the sequential activity of a PF and a nonPF. Therefore, the PFH predicts that: 1) ectopically expressed FOXA1 but not HNF4A will bind to inaccessible sites, and 2) neither TF will individually activate much tissue-specific gene expression. In order to test these predictions, we created lentiviral constructs that expressed inducible FOXA1 or HNF4A and then transduced K562 blood cells to create a FOXA1, HNF4A, and FOXA1-HNF4A double expression lines. We chose K562 cells because their epigenetic and protein binding profiles have been well characterized in published datasets, they are easy to transduce, culture, and handle, and because neither FOXA1 nor HNF4A are expressed within them.

To test the above predictions, we induced each TF and then measured expression, binding, and accessibility. The data contradicted both of the PFHs predictions. Both FOXA1 and HNF4A could bind and open inaccessible sites and both FOXA1 and HNF4A independently activated significantly enriched sets of liver- and intestine-specific genes. We further showed that HNF4A binding sites have higher target motif content than FOXA1 binding sites. This finding, in conjunction with others' data that shows that FOXA1 may bind more strongly than HNF4A (Garcia et al. 2019; Jiang, Lee, and Sladek 1997; Rufibach et al. 2006), led us to suggest that while HNF4A may have been labeled a nonPF due to its weak binding, given sufficient motif context, it can still exhibit "pioneer activity." Finally we showed that we can predict with good accuracy each TF's genome-wide binding patterns simply by using a motif count threshold. From these data we suggested that both TFs, and likely many more, can exhibit pioneer activity

and that this activity depends on TF concentration, motif content, and a TF's binding strength. These parameters suggest that pioneer activity could be captured in a quantitative metric. We published this work in January of 2022 under the title, "A test of the pioneer factor hypothesis using ectopic liver gene activation" (Hansen, Loell, and Cohen 2022).

My work in Chapter 3 follows up on the clue that pioneer activity might be quantitative. Given that pioneer activity is essentially "binding at hard to bind sites," we speculated that we could capture it by using a K_d-like metric to quantify binding strength at accessible and inaccessible genomic loci. Strong pioneer activity should lead to smaller differences in binding strength across these loci. To test the feasibility of this metric, we used our doxycycline-inducible (dox-inducible) cell lines to induce each TF across a 1,000-fold range and then measured binding at tens of thousands of sites across the genome. We called the concentration of dox at which a single site is half-maximally bound the "dox_{50}" of that site. We then compared the distribution of each TF's dox_{50}s across accessible and inaccessible loci, calculated the ratio of the mean dox_{50} of inaccessible over that of accessible loci, and called this term the TF's "Δdox_{50}." We showed that HNF4A had a smaller Δdox_{50} than FOXA1, suggesting that HNF4A has more potent pioneer activity than FOXA1. We also showed that sites at which FOXA1 targets more than 4 copies of its motif had a smaller Δdox_{50} than sites where FOXA1 targeted fewer motifs, suggesting that motif content can partially make up for weak pioneer activity. This finding is in line with our earlier discussion that pioneer activity can be summarized by whatever drives a high affinity interaction between TF and DNA.

Altogether I aimed to better understand how TFs activate silent genes. I tested the PFH, found that it did not predict the behavior of a canonical PF and nonPF, and then developed a quantitative metric for pioneer activity. I suggest that it is likely that many more TFs have some degree of pioneer activity and I encourage adoption of the Δdox_{50} or a similar metric as a way of quantifying it. Not only will these measurements add another quantitative way to explain TF behavior but will also provide us with information about how best to use TFs to ectopically activate silent genes. This information will ultimately inform rational design of future reprogramming cocktails and hopefully allow for more effective conversion processes. Then we may more fully realize the potential of establishing limitless supplies of new cells to replace those lost to damage or disease.

Chapter 2 – A Test of the Pioneer Factor Hypothesis Using Ectopic Liver Gene Activation

A test of the pioneer factor hypothesis using ectopic liver gene activation

Jeffrey L Hansen[1,2], Kaiser J Loell[1,2], Barak A Cohen[1,2*]

Affiliations

[1] The Edison Family Center for Genome Sciences and Systems Biology, School of Medicine, Washington University in St. Louis, Saint Louis, MO, USA.

[2] Department of Genetics, School of Medicine, Washington University in St. Louis, Saint Louis, MO, USA.

[*]Correspondence to: cohen@wustl.edu

This chapter was written as a paper, *A test of the pioneer factor hypothesis using ectopic liver gene activation*, with Kai Loell and Barak Cohen that was published in *eLife* (2022, January). I was the first author and designed, conducted, and analyzed most experiments. Kai Loell performed the analyses included within Figure 2.5 and Barak Cohen contributed to the design and analysis of the experiments as well as the writing of the paper.

2.1 – Abstract

The Pioneer Factor Hypothesis (PFH) states that pioneer factors (PFs) are a subclass of transcription factors (TFs) that bind to and open inaccessible sites and then recruit non-pioneer factors (nonPFs) that activate batteries of silent genes. The PFH predicts that ectopic gene activation requires the sequential activity of qualitatively different TFs. We tested the PFH by expressing the endodermal PF FOXA1 and nonPF HNF4A in K562 lymphoblast cells. While co-expression of FOXA1 and HNF4A activated a burst of endoderm-specific gene expression, we found no evidence for a functional distinction between these two TFs. When expressed independently, both TFs bound and opened inaccessible sites, activated endodermal genes, and "pioneered" for each other, although FOXA1 required fewer copies of its motif for binding. A subset of targets required both TFs, but the predominant mode of action at these targets did not conform to the sequential activity predicted by the PFH. From these results we hypothesize an alternative to the PFH where "pioneer activity" depends not on categorically different TFs but rather on the affinity of interaction between TF and DNA.

2.2 – Introduction

Transcription factors (TFs) face steric hindrance when instances of their motifs are occluded by nucleosomes(Kornberg 1974; Kaplan et al. 2009). This barrier prevents spurious transcription but must be overcome during development when TFs activate batteries of silent genes. The PFH describes how TFs recognize and activate nucleosome-occluded targets. According to the PFH, categorically different TFs cooperate sequentially to activate their targets. Pioneer factors (PFs) bind to and open inaccessible sites and then recruit non-pioneer factors (nonPFs) that are responsible for recruiting additional factors to initiate gene expression(McPherson et al. 1993; Shim, Woodcock, and Zaret 1998; L. A. Cirillo et al. 1998; Lisa Ann Cirillo et al. 2002).

PFs also play a primary role in cellular reprogramming by first engaging silent regulatory sites of ectopic lineages(Iwafuchi-Doi and Zaret 2014). Continuous overexpression of PFs and nonPFs can lead to a variety of lineage conversions(Wapinski et al. 2013; Matsuda et al. 2018; Soufi et al. 2015; Soufi, Donahue, and Zaret 2012; Sekiya and Suzuki 2011; Samantha A. Morris et al. 2014). The conversion from embryonic fibroblasts to induced endoderm progenitors offers one clear example(Sekiya and Suzuki 2011; Samantha A. Morris et al. 2014). This reprogramming cocktail combines the canonical PF FOXA1(Lisa Ann Cirillo et al. 2002) and nonPF HNF4A(Karagianni et al. 2020) and is suggested to rely upon sequential FOXA1 and then HNF4A behavior(Horisawa et al. 2020).

The PFH makes strong predictions about the activities of ectopically expressed PFs and nonPFs. Because PFs are defined by their ability to bind nucleosome-occluded instances of their motifs,

the PFH predicts that PFs should bind to a large fraction of their motifs. However, similar to other TFs, PFs only bind a limited subset of their inaccessible motifs(Barozzi et al. 2014; Mayran et al. 2018; Donaghey et al. 2018; Manandhar et al. 2017). There are chromatin states that are prohibitive to PF binding(Mayran et al. 2018; Zaret and Mango 2016) and, in at least two cases, FOXA1 requires help from other TFs to bind at its sites(Donaghey et al. 2018; Swinstead et al. 2016). These examples suggest that PFs are not always sufficient to open inaccessible chromatin. The PFH also predicts that nonPFs should only bind at accessible sites, yet the bacterial protein LexA can pioneer inaccessible sites in mammalian cells(Miller and Widom 2003). These observations, and the absence of direct genome-wide interrogations of the PFH, prompted us to design experiments to test major predictions made by the PFH using FOXA1 and HNF4A as a model PF and nonPF.

To test these predictions, we expressed FOXA1 and HNF4A separately and together in K562 lymphoblast cells and then measured their effects on DNA-binding, chromatin accessibility, and gene activation. In contrast to the predictions of the PFH, we found that both FOXA1 and HNF4A could independently bind to inaccessible instances of their motifs, induce chromatin accessibility, and activate endoderm-specific gene expression. The only notable distinction between the two factors was that HNF4A required more copies of its motif to bind. When expressed together, co-binding could only be explained in a minority of cases by sequential FOXA1 and HNF4A activity. Instead most co-bound sites required concurrent co-expression of both factors, which suggests cooperativity between these TFs at certain repressive genomic locations. We suggest that our findings present an alternative to the PFH that eliminates the categorical distinction between PFs and nonPFs and instead posits that the energy required to

pioneer occluded sites ("pioneer activity") depends on the affinity of interaction between TFs and DNA.

2.3 – Results

Generation of FOXA1 and HNF4A clonal lines

We tested predictions of the PFH using FOXA1 as a model endoderm PF and HNF4A as a model nonPF. Because PFs are defined by their behavior in ectopic settings, we expressed FOXA1 and HNF4A in mesoderm-derived K562 lymphoblast cells. These cells express neither FOXA1 nor HNF4A and present an entirely new complement of chromatin and co-factors. Thus any ectopic signature that we observe is due primarily to the TFs themselves. We focused only on the initial response to TF expression to capture primary mechanisms of TF behavior and not the secondary effects that can lead to cellular conversion and that may confound our analyses.

To perform these experiments, we created lentiviruses that inducibly express either FOXA1 or HNF4A (Figure 2.1A). We created cassettes in which a doxycycline inducible promoter drives either FOXA1 or HNF4A and cloned these cassettes separately into a lentiviral vector(Meerbrey et al. 2011) that constitutively expresses Green Fluorescent Protein (GFP). Although PFs are typically expressed at supraphysiological levels(Ng et al. 2021; Davis, Weintraub, and Lassar 1987), we infected K562 cells with each vector at a multiplicity of infection (MOI) of one to limit the degree of non-specific effects. We then used flow cytometry to sort single cells and selected FOXA1 and HNF4A clones that had similar GFP levels to ensure that our clones carried a similar transgene load. Finally, we performed both doxycycline titration induction and time

course experiments to identify the minimum doxycycline concentration and treatment time for robust TF activity. We observed that 0.5 μg/ml doxycycline for 24 hours was the minimal treatment condition that allowed *FOXA1* and *HNF4A*, and their respective target genes *ALB* and *APOB*, to reach a plateau of expression (Figure 2.S1). At this concentration, both *FOXA1 and HNF4A* were induced approximately 1,000-fold (Figure 2.S1). We used these conditions in our subsequent experiments.

Co-expression of FOXA1 and HNF4A in K562 cells conforms to the predictions of the PFH

The first prediction of the PFH is that co-expression of FOXA1 and HNF4A should be sufficient to induce ectopic tissue-specific gene expression. We tested this prediction by infecting our FOXA1 clonal line with HNF4A-expressing lentivirus to generate a double expression clonal line, hereafter referred to as FOXA1-HNF4A. Upon co-induction in K562 cells we observed strong enrichment for both liver- and intestine-specific gene activation; FOXA1-HNF4A activated 91 liver-specific genes (18 expected, $P < 10^{-38}$, cumulative hypergeometric) and 38 intestinal genes (9 expected by chance, $P < 10^{-13}$, cumulative hypergeometric) (Figure 2.1B). The dual liver and intestine enrichment that we observed is consistent with the finding that intestinal gene regulatory networks appear during reprogramming experiments that aim to use FOXA1-HNF4A to convert embryonic fibroblasts to the liver lineage(Samantha A. Morris et al. 2014). We conclude that FOXA1 and HNF4A are sufficient to activate endoderm-specific gene expression in the ectopic K562 line.

Where ectopic genes are activated in K562 cells, the PFH predicts co-binding of FOXA1 and HNF4A at inaccessible sites and induction of chromatin accessibility. Alternatively, FOXA1 and

HNF4A may not be able to overcome the K562 chromatin environment and instead activate gene expression by binding exclusively to accessible K562 sites. To distinguish between these possibilities, we measured FOXA1 and HNF4A binding by CUT&Tag(Kaya-Okur et al. 2019) after induction, and chromatin accessibility by ATAC-seq(Buenrostro et al. 2015) both before and after doxycycline induction. At the liver-specific locus *ALB*, FOXA1 and HNF4A co-bound at inaccessible sites and increased accessibility (Figure 2.1C). This pattern was consistent surrounding FOXA1-HNF4A activated liver genes: 43 of the 53 co-bound sites within 50 kb of a FOXA1-HNF4A activated gene were inaccessible prior to induction, and the accessibility signal at these co-bound sites increased substantially upon induction (Figure 2.1D-E).

Although we focused on functional binding surrounding activated liver genes, these patterns were consistent across the genome. The vast majority of both FOXA1 and HNF4A binding sites fell within sites that were inaccessible prior to induction (-dox) (Figure 2.S2) and both FOXA1 and HNF4A opened the majority of the inaccessible sites to which they bound (Figure 2.S2). These results show that despite an entirely ectopic complement of chromatin and co-factors within mesoderm-derived K562 cells, the endodermal TFs FOXA1 and HNF4A can find and activate the correct genes. Most individual binding by FOXA1 and HNF4A near their co-activated genes occurred at the same sites bound in HepG2 liver cells(Partridge et al. 2020) (Figure 2.S2). Altogether we conclude that when co-expressed, FOXA1 and HNF4A conform to the predictions of the PFH and that cis-regulatory sequences are sufficient to guide their activity within an ectopic cell type.

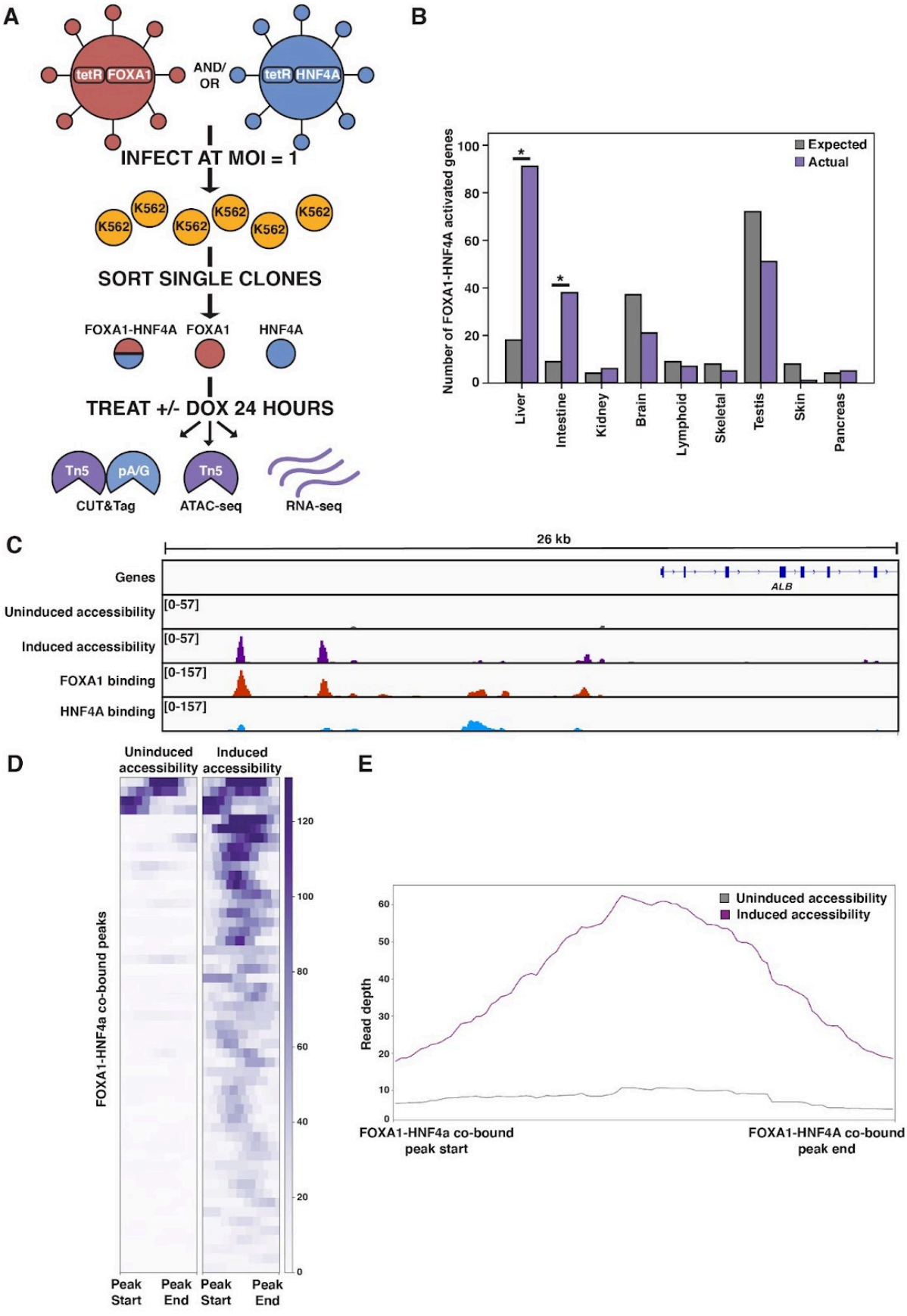
A
tetR FOXA1
AND/
OR
tetR HNF4A
INFECT AT MOI = 1
K562
SORT SINGLE CLONES
FOXA1-HNF4A
FOXA1
HNF4A
TREAT +/- DOX 24 HOURS
Tn5
pA/G
CUT&Tag
Tn5
ATAC-seq
RNA-seq
B
Number of FOXA1-HNF4A activated genes
Expected
Actual
Liver
Intestine
Kidney
Brain
Lymphoid
Skeletal
Testis
Skin
Pancreas
C
26 kb
Genes
ALB
Uninduced accessibility
Induced accessibility
FOXA1 binding
HNF4A binding
[0-57]
[0-157]
D
Uninduced accessibility
Induced accessibility
FOXA1-HNF4a co-bound peaks
Peak Start
Peak End
E
Uninduced accessibility
Induced accessibility
Read depth
FOXA1-HNF4a co-bound peak start
FOXA1-HNF4A co-bound peak end

Figure 2.1: FOXA1-HNF4A pioneers liver-specific loci in K562 cells. (A) Schematic of experimental design to infect K562 cells with FOXA1- or HNF4A-lentivirus and then perform functional assays on dox-induced cells. In CUT&Tag, a protein A-protein G fusion (pA/G) increases the binding spectrum for Fc-binding and allows Tn5 recruitment to antibody-labeled TF binding sites. In ATAC-seq, Tn5 homes to any accessible site. And in RNA-seq, polyA RNA is captured and sequenced. (B) The number of tissue-specific genes predicted from the hypergeometric distribution to be activated by FOXA1-HNF4A compared to the number actually activated. Both liver- ($P < 10\text{-}38$) and intestinal-enrichment ($P < 10\text{-}13$) are significant. There are 242 total liver-enriched genes and 122 total intestine-enriched genes. (C) Genome browser view of a representative liver-specific locus (ALB) in FOXA1-HNF4A clonal line that shows uninduced and induced accessibility, FOXA1 binding, and HNF4A binding. (D) Heatmap showing uninduced and induced accessibility at all FOXA1-HNF4A co-bound sites within 50 kb of each FOXA1-HNF4A activated liver-specific gene (n=53). (E) Meta plot showing average signal across each site from (d).

Both FOXA1 and HNF4A individually activate many liver-specific genes

We next sought to test whether ectopic tissue-specific gene expression in K562 cells results from the sequential activity of FOXA1 and HNF4A, as predicted by the PFH. Sequential activity predicts that HNF4A will not bind to its sites without FOXA1, and that FOXA1 won't activate expression without HNF4A, such that neither FOXA1 nor HNF4A should activate tissue-specific gene expression when expressed alone. To test this prediction, we used the single expression K562 lines to induce either FOXA1 or HNF4A alone and measured mRNA expression by RNA-seq. FOXA1 induction resulted in strong liver-specific enrichment ($P < 10^{-4}$, cumulative Hypergeometric) and weak intestinal-specific enrichment (not significant) (Figure 2.2A), while HNF4A induction resulted in both strong liver-specific enrichment ($P < 10^{-8}$, cumulative Hypergeometric) and strong intestinal-specific enrichment ($P < 10^{-15}$, cumulative Hypergeometric) (Figure 2.2B). Importantly, neither FOXA1 nor HNF4A are expressed within K562 cells nor did they induce expression of the other TF, suggesting that the expression changes we observed were due to the independent effects of either FOXA1 or HNF4A.

When expressed individually, FOXA1 and HNF4A activated largely independent sets of liver genes (Figure 2.2C) and intestinal genes (Figure 2.2D). FOXA1 activates liver genes enriched

for fibrinolysis and complement activation (Table 2.S1) whereas HNF4A activates liver genes enriched for cholesterol import and lipoprotein remodeling (Table 2.S2). Thus, in contrast to the predictions of the PFH, FOXA1 and HNF4A are each sufficient to induce separate and specific endodermal responses when expressed alone in K562 cells.

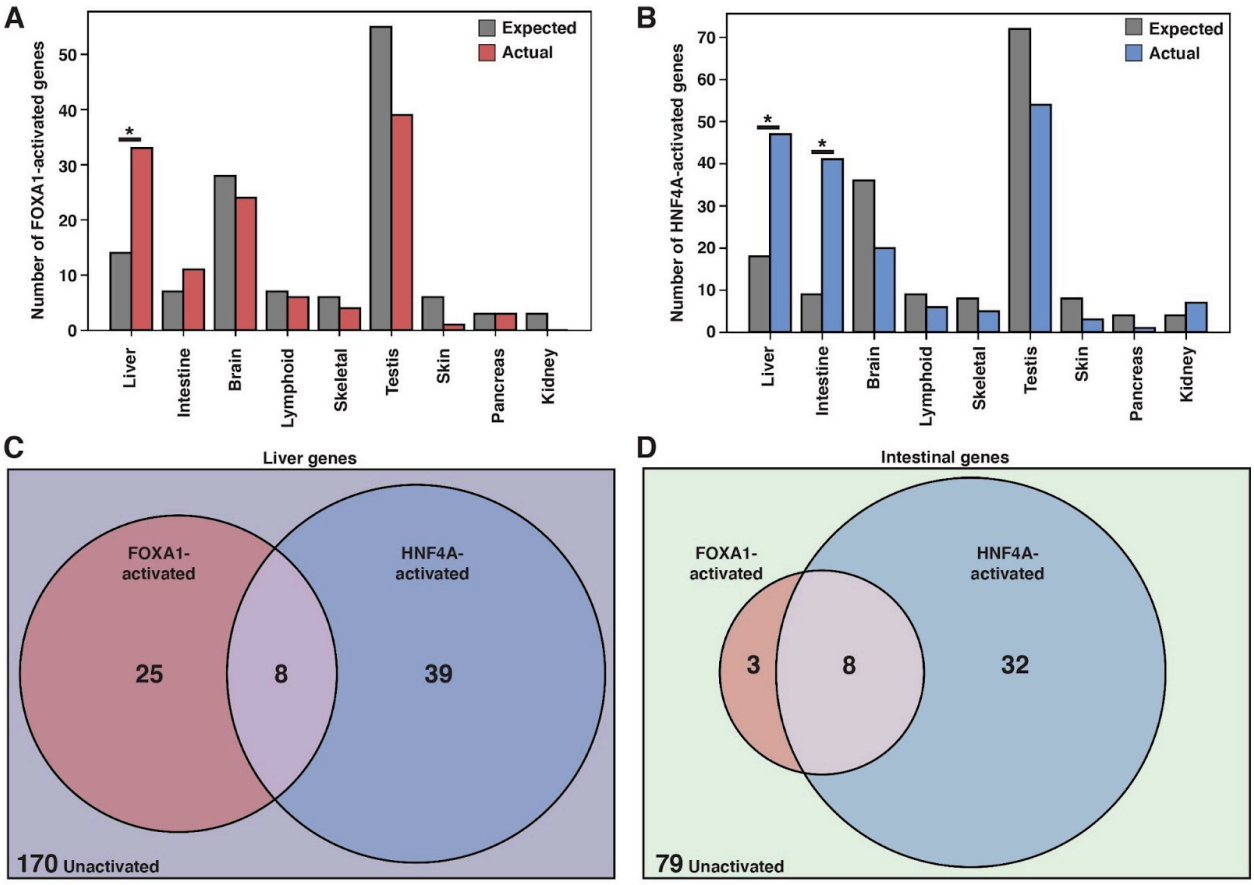

Figure 2.2: FOXA1 and HNF4A activate independent liver- and intestine-specific genes. (A) The number of tissue-specific genes predicted from the hypergeometric distribution to be activated by FOXA1 compared to the number actually activated. Liver-enrichment ($P < 10\text{-}4$) is significant. There are 242 total liver-enriched genes. **(B)** The number of tissue-specific genes predicted from the hypergeometric distribution to be activated by HNF4A compared to the number actually activated. Liver- ($P < 10\text{-}8$) and intestine-enrichment ($P < 10\text{-}15$) are significant. There are 242 total liver-enriched genes and 122 total intestine-enriched genes. **(C)** 242 liver genes characterized as activated by Foxa1, HNF4A, both, or neither. **(D)** 122 intestine genes characterized as activated by FOXA1, HNF4A, both, or neither.

Both FOXA1 and HNF4A can independently bind and open inaccessible sites around liver genes

Our results raised the possibility that both FOXA1 and HNF4A can bind and open inaccessible instances of their motifs. To test this, we induced FOXA1 and HNF4A expression individually and then measured each factor's binding profile and their accessibility profiles before and after induction. FOXA1 induction resulted in FOXA1 binding and induced accessibility adjacent to *ARG1,* a liver-specific gene that is silent in K562 cells (Figure 2.3A), while HNF4A alone bound and induced accessibility at sites nearby the liver-specific gene *APOC3* (Figure 2.3B). This pattern was consistent across liver-specific loci. 34 of the 59 FOXA1 binding sites within 50 kb of a FOXA1-activated liver gene were inaccessible and opened upon induction (Figure 2.3C,E) as was the case for 39 of the 76 HNF4A binding sites (Figure 2.3D,F). We observed similar patterns genome-wide. FOXA1 and HNF4A bound primarily to sites that were inaccessible prior to induction (-dox) (Figure 2.S3), opened them (Figure 2.S3), and in regions surrounding activated genes, most binding occurred at the same sites bound in HepG2 liver cells (Figure 2.S3). We conclude that FOXA1 and HNF4A have roughly equivalent abilities to bind and open inaccessible sites.

We sought to reconcile these findings with what the PFH had predicted. We first considered whether, in the absence of FOXA1, native K562 TFs were "pioneering" for HNF4A. A *de novo* motif discovery analysis of the 500 bp centered on inaccessible FOXA1 or HNF4A binding sites revealed strong enrichment for each TF's motif, but no other strong signals. Similarly, we found no evidence for enrichment of predicted K562 PFs AP1 (FOS/JUN) (MA0099.2) (Biddie et al. 2011), GATA1 (MA0035.4) (Iwafuchi-Doi and Zaret 2014), MYB (MA0100.1) (Lemma et al.

2021), or SPI1 (PU.1) (MA0080.1) (Iwafuchi-Doi and Zaret 2014), either in inaccessible binding sites over randomly chosen sites, or in HNF4A over FOXA1 binding sites (Figure 2.S4). Thus, the similar activities of FOXA1 and HNF4A are not explained by pioneering activity provided by endogenous K562 TFs.

We next considered whether differences in FOXA1 and HNF4A motif content could explain our results. We focused on binding sites surrounding activated liver genes and used FOXA1 and HNF4A position weight matrices (Figure 2.3G) to count occurrences in the 500 bp of sequence surrounding these sites. Sites independently pioneered by FOXA1 contained between 2-4 motifs, while sites pioneered by HNF4A contained 3-6 motifs (Figure 2.3H). This is despite the fact that the FOXA1 motif occurs more frequently across the genome than the HNF4A motif (Figure 2.S5). This observation is consistent with data showing that FOXA1 has higher affinity for its binding site than HNF4A(Garcia et al. 2019; Rufibach et al. 2006; Jiang, Lee, and Sladek 1997) and suggests that there may not be anything categorically different about FOXA1 and HNF4A, but rather that "pioneer activity" may depend on the affinity of interaction between TF and DNA.

Another possible explanation for our results could be that at the concentrations TFs are expressed in cellular reprogramming, the differences between PFs and nonPFs are no longer apparent. We took advantage of our doxycycline-inducible system to test this hypothesis by lowering the doxycycline concentration from 0.5 µg/ml to 0.05 µg/ml, thus dropping the TF concentration significantly (Figure 2.S1). We then re-measured binding and expression. We found that lower induction resulted in far fewer FOXA1 and HNF4A genome-wide binding events (Figure 2.S6).

This effect was even more pronounced when we subset the binding events into sites that were either accessible or inaccessible prior to induction. Both FOXA1 and HNF4A shifted from binding predominantly inaccessible sites to binding predominantly accessible sites (Figure 2.S6). Thus binding of both factors depends on a balance of TF concentration and accessibility state, and the results from expression profiling in the lower induction regime are consistent with this idea. Whereas FOXA1 and HNF4A previously activated 33 and 47 liver genes, at the lower induction rate they activated 8 and 30, respectively (Figure 2.S6). Thus, lowering the induction levels had strong effects on the activities of both FOXA1 and HNF4A, but did not reveal qualitative differences between the two TFs. These results suggest that the induction conditions in cellular reprogramming do not mask differences between the TFs, a result consistent with the fact that the PFH was developed to explain the properties of cellular reprogramming cocktails.

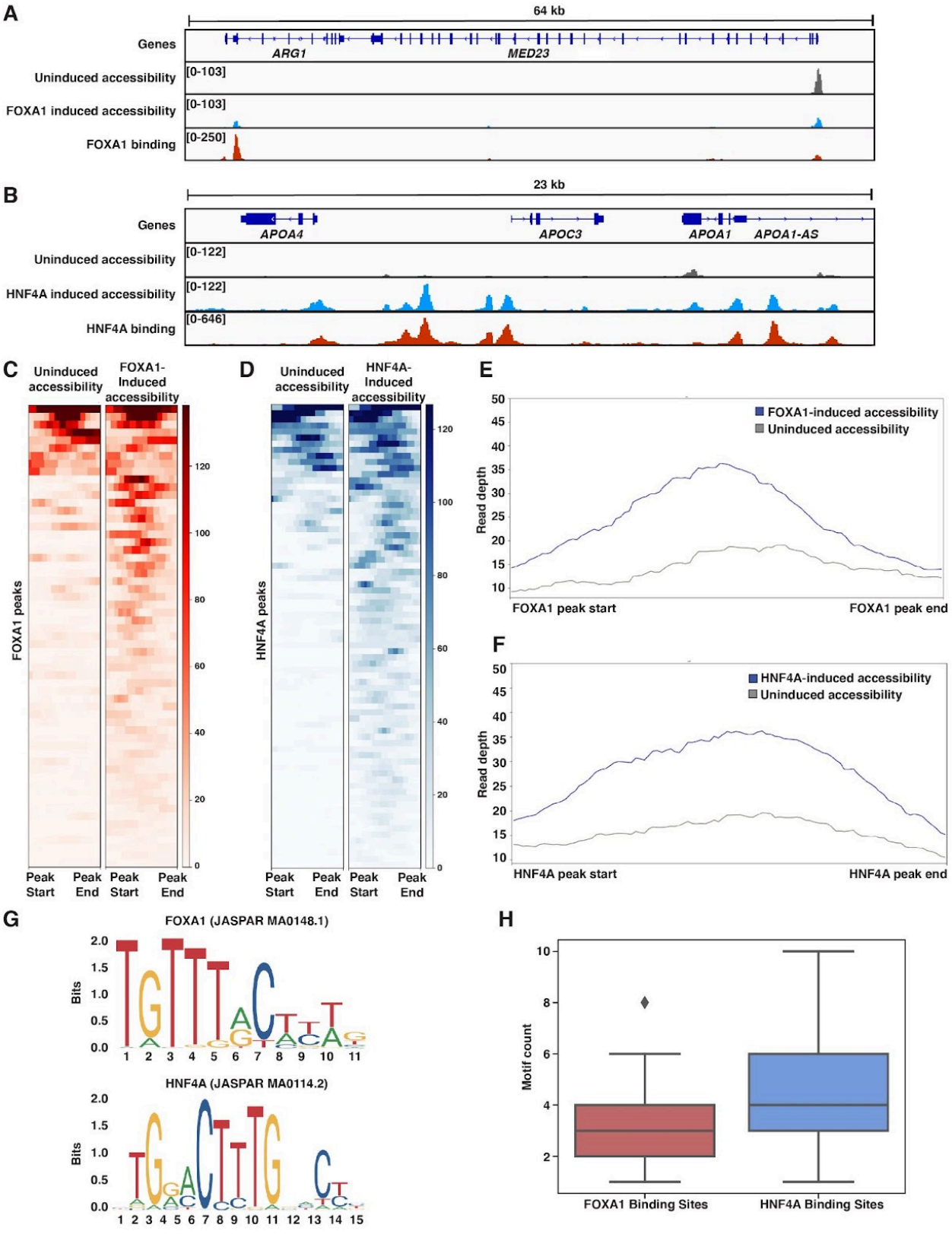
A
64 kb
Genes
ARG1
MED23
Uninduced accessibility
[0-103]
FOXA1 induced accessibility
[0-103]
FOXA1 binding
[0-250]
B
23 kb
Genes
APOA4
APOC3
APOA1
APOA1-AS
Uninduced accessibility
[0-122]
HNF4A induced accessibility
[0-122]
HNF4A binding
[0-646]
C
Uninduced accessibility
FOXA1-Induced accessibility
FOXA1 peaks
Peak Start
Peak End
Peak Start
Peak End
D
Uninduced accessibility
HNF4A-Induced accessibility
HNF4A peaks
Peak Start
Peak End
Peak Start
Peak End
E
FOXA1-induced accessibility
Uninduced accessibility
Read depth
FOXA1 peak start
FOXA1 peak end
F
HNF4A-induced accessibility
Uninduced accessibility
Read depth
HNF4A peak start
HNF4A peak end
G
FOXA1 (JASPAR MA0148.1)
Bits
HNF4A (JASPAR MA0114.2)
Bits
H
Motif count
FOXA1 Binding Sites
HNF4A Binding Sites

Figure 2.3: Both FOXA1 and HNF4A can pioneer liver-specific loci. (A) Genome browser view of a representative liver-specific locus (ARG1) in FOXA1 clonal line showing uninduced and induced accessibility and FOXA1 binding. **(B)** Genome browser view of a representative liver-specific locus (APOC3) in HNF4A clonal line showing uninduced and induced accessibility and HNF4A binding. **(C)** Heatmap of uninduced and induced accessibility at all FOXA1 binding sites within 50 kb of each FOXA1-activated liver-specific genes (n = 59). **(D)** Heatmap of uninduced and induced accessibility at all HNF4A binding sites within 50 kb of each HNF4A-activated liver-specific genes (n = 76). **(E)** Meta plot showing average signal across each site from (c). **(F)** Meta plot showing average signal across each site from (d). **(G)** Human FOXA1 and HNF4A sequence logo from JASPAR. **(H)** FOXA1 or HNF4A motif count within 500 bp centered upon FOXA1 or HNF4A binding sites within 50 kb of each FOXA1- or HNF4A-activated liver-specific genes, respectively. Motifs were called with FIMO using 1e-3 a p-value threshold. For each boxplot, the center line represents the median, the box represents the first to third quartiles, and the whiskers represent any points within 1.5 times the interquartile range.

Some liver genes require cooperative FOXA1-HNF4A activity

In addition to those genes independently activated by FOXA1 and HNF4A, there is an additional set of 31 liver genes that are not activated until both FOXA1 and HNF4A are present (Figure 2.4A). We therefore asked whether these 31 liver genes are activated sequentially, as predicted by the PFH. If these genes conform to the PFH, then we would expect that at every gene, there are nearby sites where FOXA1 binds individually and where FOXA1 and HNF4A co-bind when expressed together. This would be evidence for FOXA1 "pioneering" sites for later HNF4A binding and so we have called these sites "FOXA1 Pioneered" (FP). Sites are "HNF4A Pioneered" (HP) if HNF4A binds individually and FOXA1 and HNF4A co-bind when expressed together and sites are "Cooperatively Bound" (CB) if neither TF binds individually but both do when expressed together.

When there is sequential binding of the two TFs it will be apparent in comparisons of the single versus double expression clones, whereas obligate cooperativity between the TFs will result in binding that is observed only in the double expression clone. There are examples of each modality surrounding *AMDHD1*, a liver-specific gene co-activated by FOXA1 and HNF4A (Figure 2.4B). When we examine all of the liver genes only activated by FOXA1-HNF4A co-

expression, we find that in contradiction with the PFH, there are roughly equal numbers of FP, HP, and CB sites (Figure 2.4C). Therefore, in most cases, genes that require joint FOXA1-HNF4A activity do not rely on sequential FOXA1-then-HNF4A behavior.

The patterns of genome-wide co-binding and accessibility of FOXA1 and HNF4A follow similar trends. Of the 11,402 co-bound sites, 2,023 were FP, 3,398 were HP, and 2,192 were CB (Figure 2.4D) and FOXA1-induced differentially accessible peaks explain a minority of the FOXA1-HNF4A differentially accessible peaks (Figure 2.S7). Cooperative binding may be more important in less accessible parts of the region, as there are more CB sites in ChromHMM-labeled(Ernst and Kellis 2012) heterochromatic and repressed regions, and there are more FP and HP sites in promoter and enhancer regions (Figure 2.4E).

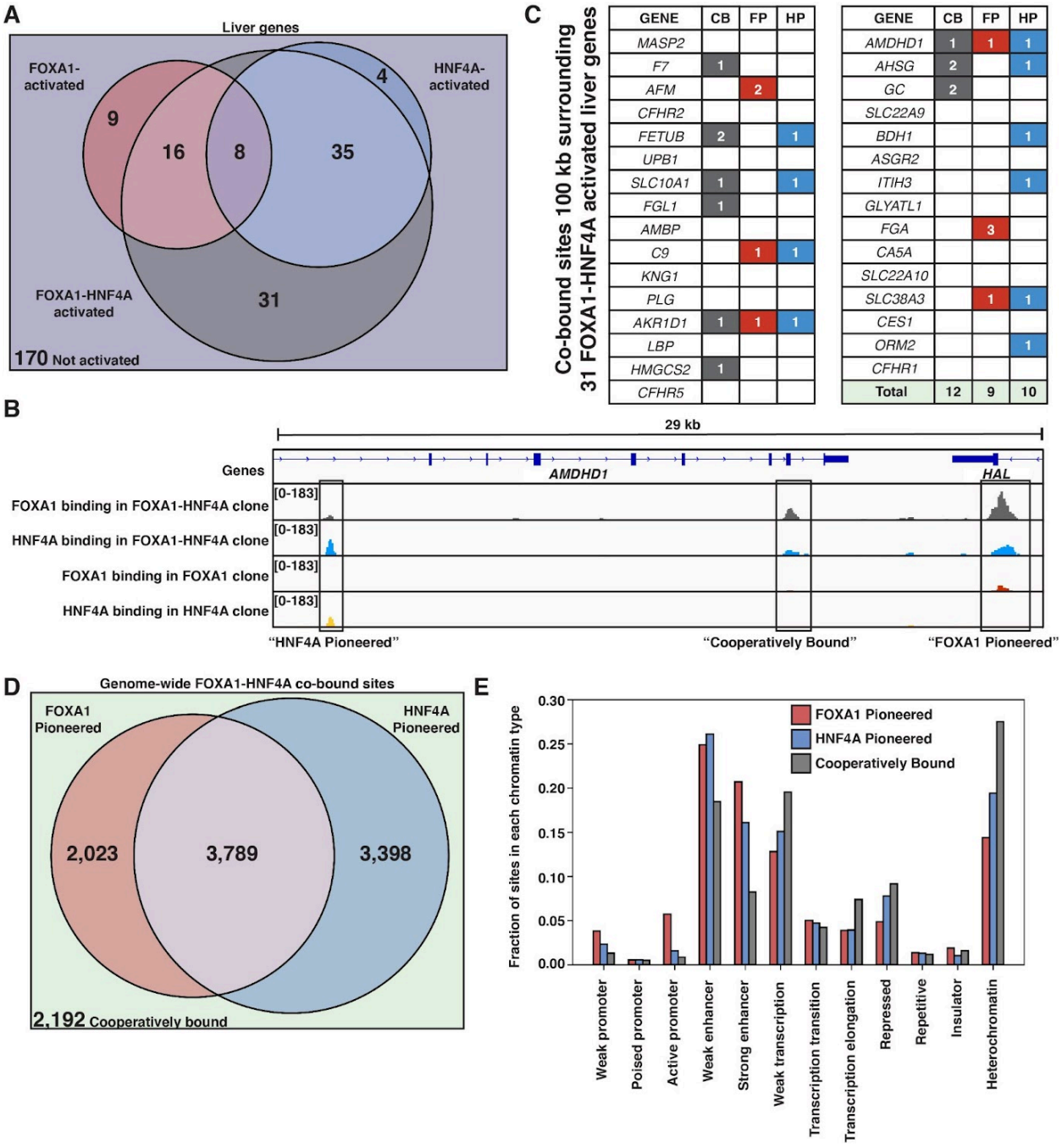

Figure 2.4: FOXA1 and HNF4A both pioneer and cooperate at liver-specific sites. (A) Venn diagram of all liver genes categorized as either activated by FOXA1, HNF4A, FOXA1-HNF4A, some combination, or by none of the three cocktails. **(B)** Genome browser view of a representative liver-specific locus (*AMDHD1*) showing examples of a co-bound site that is "FOXA1 Pioneered" (FP), "HNF4A Pioneered" (HP), and "Cooperatively Bound" (CB). The first two tracks are FOXA1 and HNF4A binding in the FOXA1-HNF4A co-expression clone and the last two tracks are FOXA1 and HNF4A binding in their individual expression clones. **(C)** List of the 31 liver genes that are only

activated by FOXA1-HNF4A co-expression. The columns indicate how many co-bound FP, HP, or CB peaks exist within 100 kb of the gene. **(D)** Venn diagram of all genome-wide co-bound peaks categorized as either bound by FOXA1 individually (FP), HNF4A individually (HP), by both, or by neither (CB). **(E)** Overlap of FP, HP, and CB sites from (D) with ChromHMM annotations showing the fraction of each co-binding site type in each chromatin region.

Genome-wide motif analysis supports affinity model

The correlation between TF binding and factors such as TF binding strength, motif content, TF concentration, and accessibility state have so far suggested that an affinity model may explain ectopic FOXA1 and HNF4A behavior. Thus we predicted that motif counts would explain genome-wide binding patterns. Because it requires more energy to bind at inaccessible sites than accessible sites, we predicted that there would be more motifs at inaccessible binding sites than at accessible sites, and that this motif distribution would be higher than that found in random genomic sequences. When we examined the 500 bp of sequence centered upon genome-wide TF binding sites, we found that for both FOXA1 and HNF4A, inaccessible binding sites had higher motif content than accessible binding sites and that these inaccessible binding sites had higher motif content than random inaccessible regions (Figure 2.5A-B). A simple motif threshold could predict binding , though only when predicting inaccessible sites (Figure 2.5C).

We also predicted that if FOXA1 and HNF4A are not categorically different, then we would find similar trends between the motifs for the two TFs . We predicted that total FOXA1 and HNF4A motif count at inaccessible sites would be higher than at random sites, and that FP or HP sites would have more FOXA1 or HNF4A sites, respectively, than CB sites. When we examined the 500 bp of sequence centered upon genome-wide co-bound sites, we found that there was higher total motif content at inaccessible binding sites as compared to random (Figure 2.5D) and that FOXA1 and HNF4A motif content was higher at FP or HP sites, respectively, than CB sites

(Figure 2.5E). And like individually bound sites, a motif threshold could only predict inaccessible binding behavior (Figure 2.5F, top panels). The motif threshold was somewhat effective at differentiating between FP or HP versus CB sites (Figure 2.5F, lower panel). Altogether these results further support our hypothesis that affinity better explains ectopic FOXA1 and HNF4a “pioneer activity” than the current formulation of the PFH.

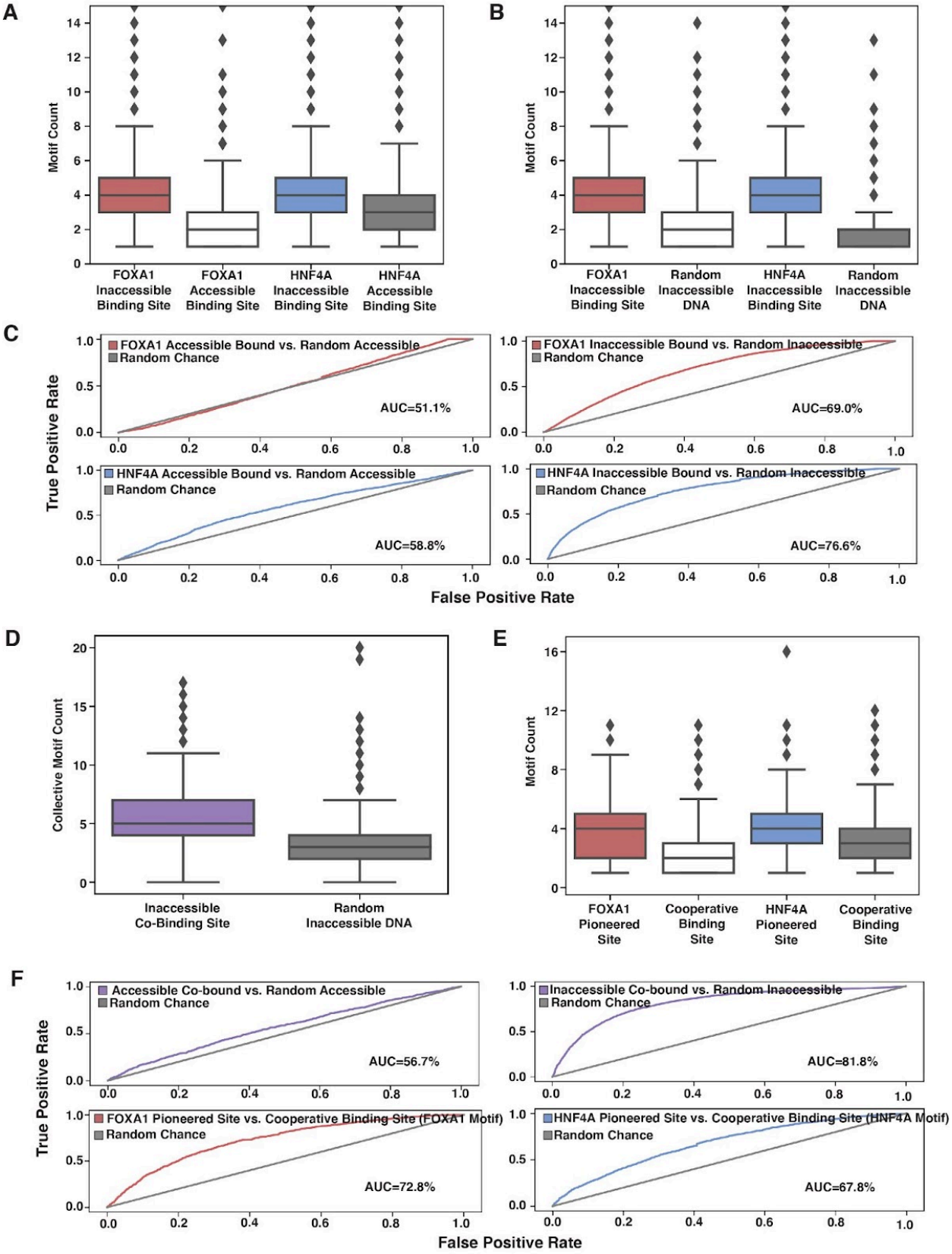
A
Motif Count
FOXA1 Inaccessible Binding Site
FOXA1 Accessible Binding Site
HNF4A Inaccessible Binding Site
HNF4A Accessible Binding Site
B
Motif Count
FOXA1 Inaccessible Binding Site
Random Inaccessible DNA
HNF4A Inaccessible Binding Site
Random Inaccessible DNA
C
True Positive Rate
FOXA1 Accessible Bound vs. Random Accessible
Random Chance
AUC=51.1%
FOXA1 Inaccessible Bound vs. Random Inaccessible
Random Chance
AUC=69.0%
HNF4A Accessible Bound vs. Random Accessible
Random Chance
AUC=58.8%
HNF4A Inaccessible Bound vs. Random Inaccessible
Random Chance
AUC=76.6%
False Positive Rate
D
Collective Motif Count
Inaccessible Co-Binding Site
Random Inaccessible DNA
E
Motif Count
FOXA1 Pioneered Site
Cooperative Binding Site
HNF4A Pioneered Site
Cooperative Binding Site
F
True Positive Rate
Accessible Co-bound vs. Random Accessible
Random Chance
AUC=56.7%
Inaccessible Co-bound vs. Random Inaccessible
Random Chance
AUC=81.8%
FOXA1 Pioneered Site vs. Cooperative Binding Site (FOXA1 Motif)
Random Chance
AUC=72.8%
HNF4A Pioneered Site vs. Cooperative Binding Site (HNF4A Motif)
Random Chance
AUC=67.8%
False Positive Rate

Figure 2.5: Affinity model predicts binding events. (A) FOXA1 or HNF4A motif count at all genomic occurrences of the respective TF's accessible or inaccessible binding sites. **(B)** FOXA1 or HNF4A motif count in genome-wide inaccessible binding sites versus length-matched random inaccessible DNA sequences. **(C)** Receiver operating characteristic (ROC) curves for predictive power of using sequence motif content to predict accessible (left panels) or inaccessible (right panels) binding sites from random sequence. **(D)** Total FOXA1 and HNF4A motif count at all genomic occurrences of inaccessible co-binding versus length-matched random inaccessible DNA sequences. **(E)** FOXA1 or HNF4A motif count in respective FOXA1 or HNF4A Pioneered sites versus in cooperative binding sites (where neither TF bound individually). **(F)** ROC curves for predictive power of using sequence motif content to predict accessible or inaccessible co-binding events from random sequence (top panels) or to predict FOXA1 or HNF4A pioneered events from cooperative binding events. All FIMO scans used 1e-3 as p-value threshold and were conducted on 500bp of sequence centered upon the binding site.

2.4 – Discussion

In contrast to the predictions of the PFH, we found that both the canonical PF FOXA1 and nonPF HNF4A can independently bind inaccessible sites, increase accessibility, and activate nearby endodermal genes. Some endodermal genes require the joint activity of both TFs, but the predominant mode of action at these targets does not conform to the predicted sequential activity of FOXA1 followed by HNF4A. These observations suggest that we do not need to invoke the PFH to explain FOXA1 and HNF4A's behavior in ectopic K562 cells and that instead we may use the affinity of interaction between each TF and its target sites to explain its behavior.

An affinity model assumes that there is nothing categorically different between FOXA1 and HNF4A. We hypothesize that differences still exist between TFs' abilities to bind at nucleosome-occluded sites but that "pioneer activity" is a spectrum not a binary classifier. The probability of a binding event depends on the intrinsic binding ability of the TF and the motif count at a potential binding site. Previous measures of intrinsic binding strength that show FOXA1 binds more tightly than HNF4A (Garcia et al. 2019; Rufibach et al. 2006; Jiang, Lee, and Sladek 1997) may explain why in our assays FOXA1 requires fewer copies of its motif to bind. In fact FOXA1

has a special three-dimensional, histone-like structure that may explain its superior binding strength(Clark et al. 1993).

However, given the right sequence context, HNF4A also displays pioneer activity. We hypothesize that HNF4A was mis-classified because of both developmental timing and indirect assays of pioneer activity. FOXA1 precedes HNF4A during hepatic development (Lau et al. 2018) and studies have traditionally established pioneer factor status by using endogenous binding or genome-wide chromatin marks. Perhaps sequential activity of FOXA1 and HNF4A is necessary during hepatic development, but our data show that both TFs are sufficient to independently activate silent genes.

We further hypothesize that our findings may extend to other reprogramming cocktails that combine PFs and nonPFs. While our study is limited to two TFs at two concentrations in one cell line, other data support our hypothesis. Early reprogramming of fibroblasts to myoblasts relied solely upon the ectopic overexpression of MyoD, without an accompanying nonPF(Davis, Weintraub, and Lassar 1987; Choi et al. 1990) and new reprogramming cocktails have been tested and validated in a large-scale screen for single, cell autonomous reprogramming TFs(Ng et al. 2021). Increasing the efficiency of reprogramming cocktails that depend on multiple TFs will require distinguishing between the independent and cooperative effects of TFs. For example, our finding that HNF4A independently activates more intestine-specific genes than FOXA1 raises the possibility that titrating down HNF4A activity during reprogramming could result in a more liver-specific profile. Such fine-tuning of TF activities has been suggested as an option to

improve the success of other reprogramming cocktails(Ma et al. 2015; Wang et al. 2015; Vaseghi et al. 2016).

Although we found clear instances of sites independently pioneered by either FOXA1 or HNF4A, not all sites containing multiple motifs were pioneered in K562 cells, which comports with studies showing that the sequence context in which motifs occur also plays an important role in determining whether sites will be pioneered or not. GAL4's ability to bind nucleosomal DNA templates depends both on the number of copies of its motif(Workman, Schuetz, and Kingston 1991) and the positioning of the motif in the nucleosome(Vettese-Dadey et al. 1994). Precise nucleosome positioning also dictates TP53 and OCT4 pioneering behavior(Yu and Buck 2019; Huertas et al. 2020). A TF's motif affinity, motif count, and the presence of co-factor motifs are all strong predictors of pioneer activity(Yan, Chen, and Bai 2018; Manandhar et al. 2017; Donaghey et al. 2018; Heinz et al. 2010; Boyes and Felsenfeld 1996; Minderjahn et al. 2020; Meers, Janssens, and Henikoff 2019) and certain types of heterochromatic patterning have been labeled "pioneer resistant"(Mayran et al. 2018). Thus we hypothesize that general pioneer activity may best be summarized by the free energy balance between TFs, nucleosomes, and DNA(Polach and Widom 1996; Mirny 2010) rather than as a property of specific classes of TFs**.**

2.5 – Materials and Methods

Cell lines

We grew K562 cells (ATCC CCL-243, Manassas, VA) in Iscove's Modified Dulbecco Serum supplemented with 10% fetal bovine serum, 1% penicillin-streptomycin and 1% non-essential

amino acids. We used these cells to generate our clonal lines (FOXA1, HNF4A, and FOXA1-HNF4A) and we thank the Washington University in St. Louis Genome Engineering and iPSC Center for their help confirming K562 identity with STR profiling and testing for mycoplasma contamination. When it was time to conduct one of our functional assays, we split FOXA1-, HNF4A-, or FOXA1-HNF4A-expressing cells into replicate flasks and then treated with either +/- 0.5 μg/ml or 0.05 μg/ml doxycycline (Sigma #D9891-1G) for 24 hours.

Cloning, production, and infection of viral vectors

We used PCR to add V5 epitope tags to the 3' end of FOXA1 (Addgene #120438, Watertown, MA) and HNF4A (Addgene #120450) constructs and then used HiFi DNA Assembly (NEB #E2621L, Ipswich, MA) to clone each construct into a pINDUCER21 doxycycline-inducible lentiviral vector (Addgene #46948). All primers are listed in Supplementary file 1. The Hope Center Viral Vector Core at Washington University in St. Louis then generated and titered high-concentration virus. We infected human K562 cells at a multiplicity of infection of 1 by spinoculation at 800G for 30 minutes in the presence of 10 μg/ml polybrene (Sigma #TR1003G, St. Louis, MO), passaged the cells for 3 days, and then selected for positively-infected cells by single cell sorting on GFP+ into 96-well plates. Finally we used qPCR to select for clones that had high inducibility of TF and target gene expression (Figure 2.S1).

RNA extractions, reverse transcription, and qPCR

We extracted RNA from 1e6 cells/sample with the PureLink RNA Mini (Invitrogen #12183020, Waltham, MA) column extraction kit and completed on-column DNA digestion with PureLink DNase (Invitrogen #12185010). We quantified and assessed the quality of the RNA with an

Agilent 2200 Tapestation instrument and then either froze down pure RNA for later RNA-sequencing library preparation or used ReadyScript cDNA Synthesis Mix (Sigma #RDRT-100RXN) to produce cDNA for qPCR. We performed qPCR with SYBR Green PCR Master Mix (Applied Biosystems #4301955, Waltham, MA) and gene-specific and housekeeping primers (Supplementary File 1).

RNA-sequencing and analysis

We generated three replicates of +/- doxycycline-treated RNA-sequencing libraries with the NEBNext Ultra II Directional RNA Library Prep Kit (NEB #E7765S). We quantified and assessed the quality of the libraries with an Agilent 2200 Tapestation instrument, size selected with AMPure XP beads (Beckman Coulter #A63880, Brea, CA), and then sequenced the libraries with 75bp paired-end reads on an Illumina NextSeq 500 instrument.

We quantified transcripts with Salmon(Patro et al. 2017), filtered out any with fewer than 10 reads, and then called differentially expressed transcripts with DESeq2(Love, Huber, and Anders 2014). A gene was called differentially upregulated if it had a log2fold change of at least 1 and was called “activated” if it had fewer than 50 normalized reads in the uninduced control. A gene was called “tissue-specific” according to the Human Protein Atlas definition of tissue enrichment(Uhlén et al. 2015), which is if a gene is at least 4-fold higher expressed in the tissue-of-interest than in any other tissue as measured by deep sequencing of RNA from the tissue-of-interest.

ATAC-sequencing and analysis

We followed the Omni-ATAC protocol(Ryan Corces et al. 2017) to generate two replicates of +/- doxycycline-treated low-background ATAC-sequencing libraries. We isolated 2e5 cells/sample and then extracted 5e4 nuclei/sample for tagmentation and library preparation. We quantified and assessed the quality of the libraries with an Agilent 2200 Tapestation instrument, size selected with AMPure XP beads, and then sequenced the libraries with 75bp paired-end reads on an Illumina NextSeq 500 instrument.

We aligned transcripts with bowtie2(Langmead and Salzberg 2012) with the parameters: --local -X2000, generated RPKM normalized BigWig files for visualization with DeepTools bamCoverage(Ramírez et al. 2016), and then called peaks at low stringency with macs2 ($p = 0.01$)(Y. Zhang et al. 2008). With these peaks, we either called reproducible peaks with IDR (FDR of 0.05)(Li et al. 2011) or used DiffBind(Stark, Brown, and Others 2011) to call differential peaks. We calculated the Fraction of Reads in Peaks (FRiP) with the Subread featureCounts tool (Liao, Smyth, and Shi 2014), counting reads for each replicate in the IDR-merged peak list (Table 2.S3).

CUT&Tag and analysis

We followed the CUTANA Direct-to-PCR CUT&Tag protocol (EpiCypher, Chapel Hill, NC) to generate two replicates of low-background CUT&Tag libraries. We isolated 1e5 cells/sample, extracted nuclei with Concanavalin A paramagnetic beads (Epicypher #21-1401), and then either used rabbit anti-human FOXA1 monoclonal antibody (Cell Signaling #53528, Danvers, MA), mouse anti-human HNF4A monoclonal antibody (Invitrogen #MA1-199), or rabbit anti-human histone H3K4me3 polyclonal antibody (Epicypher #13-0041) as a positive control. We amplified

this signal with either goat anti-rabbit (Epicypher #13-0047) or goat anti-mouse (Epicypher #13-0048) polyclonal secondary antibodies. For a negative control, we omitted the primary antibody and checked for any non-specific pull-down. Finally, we used CUTANA pAG-Tn5 (Epicypher #15-1017) to tagment the genomic regions surrounding each bound antibody complex. We quantified and assessed the quality of the libraries with an Agilent 2200 Tapestation instrument, size selected with AMPure XP beads, and then sequenced the libraries with 150bp paired-end reads on an Illumina NextSeq 500 instrument.

When we assessed our libraries with the Agilent Tapestation instrument, we found that our negative controls had minimal signal. This is expected in the protocol and as such sequencing the sample is recommended as optional(Kaya-Okur et al. 2020). For this reason, we sequenced only our positive samples. We aligned our samples with Bowtie2(Langmead and Salzberg 2012) using recommended parameters(Kaya-Okur et al. 2020): --very-sensitive --end-to-end --no-mixed --no-discordant -I 10 -X700, created RPKM normalized BigWig files with DeepTools bamCoverage(Ramírez et al. 2016), and called peaks with macs2 (p = 1e-5)(Y. Zhang et al. 2008) with recommended parameters(Kaya-Okur et al. 2019). We calculated the Fraction of Reads in Peaks (FRiP) with Subread featureCounts tool (Liao, Smyth, and Shi 2014) (Table 2.S4). We then combined overlapping peaks from replicate samples using BEDTools intersect(Quinlan and Hall 2010). We attributed binding sites to genes if they were within 50 kb (25 kb up- and 25 kb downstream) of the gene's TSS. Because co-binding occurred less frequently, we attributed co-binding sites to genes if they were within 100 kb of the gene's TSS. "FOXA1 Pioneered" sites were those where we identified overlapping FOXA1 and HNF4A binding peaks within 100 kb of a gene that was only activated by FOXA1 and HNF4A and

where there was also an overlapping FOXA1 binding peak, when FOXA1 was expressed alone. "HNF4A Pioneered" sites were those where we identified overlapping FOXA1 and HNF4A binding peaks within 100 kb of a gene that was only activated by FOXA1 and HNF4A and where was also an overlapping HNF4A binding peak, when HNF4A was expressed alone. And "Cooperatively Bound" sites were those where we identified overlapping FOXA1 and HNF4A binding peaks within 100 kb of a gene that was only activated by FOXA1 and HNF4A and where there was neither a FOXA1 nor HNF4A binding peak.

Tissue- and biological process-specific expression analysis

We generated lists of tissue-specific genes for each tissue by extracting "enriched genes" from the Human Protein Atlas, as detailed above. We then computed hypergeometric assays to determine if our activated genes were enriched in any tissue-specific gene set. Finally, we used Panther gene ontology analysis to identify enriched biological processes.

Genome tracks and profile plot analysis

We visualized the signal from our functional assays by loading each file into the Integrated Genome Viewer(Robinson et al. 2011), using hg19 as reference. We then used the computeMatrix function in reference-point mode and plotProfile function, both with default parameters, in the DeepTools suite(Ramírez et al. 2016) to display aggregated CUT&Tag and ATAC-sequencing signals across indicated genomic regions.

Motif and chromatin segmentation analysis

Before running motif scans, we extracted 500bp of sequence centered on the binding sites of interest. Then we used STREME (Bailey 2021) for de novo motif discovery and FIMO (Grant, Bailey, and Noble 2011) for specific motif occurrence counting. We used 1e-3 as a p-value threshold and JASPAR (Fornes et al. 2020) PWMs for FOXA1 (MA0148.1) and HNF4A (MA0114.2). To use motif content to predict binding, we lowered the p-value threshold to 0 to allow for weak motif contributions and then summed the motif content for each sequence. A simple threshold on this aggregate score was used as a classifier, with the ROC curves generated by sweeping this threshold and plotting the resulting true positive rates against false positive rates. We used ChromHMM annotations(Ernst and Kellis 2012) to characterize the epigenetic profile of FOXA1 and HNF4A binding sites.

Data Availability

All genomic sequencing data have been deposited on Gene Expression Omnibus (GEO) under accession number GSE182191.

2.6 – Acknowledgements

We thank Dr. Gary Stormo, Dr. Robi Mitra, and members of the Cohen Lab for reading and critiquing the manuscript and for helpful discussion; Jessica Hoisington-Lopez and MariaLynn Crosby in the DNA Sequencing Innovation Lab for assistance with high-throughput sequencing; the Genome Engineering and iPSC Center for allowing us to use their Sony Flow Cytometer for cell sorting; and Mingjie Li in the Hope Center Viral Vectors Core for assistance with producing lentiviral expression vectors.

2.7 – Supplementary Information

Supplementary Figures

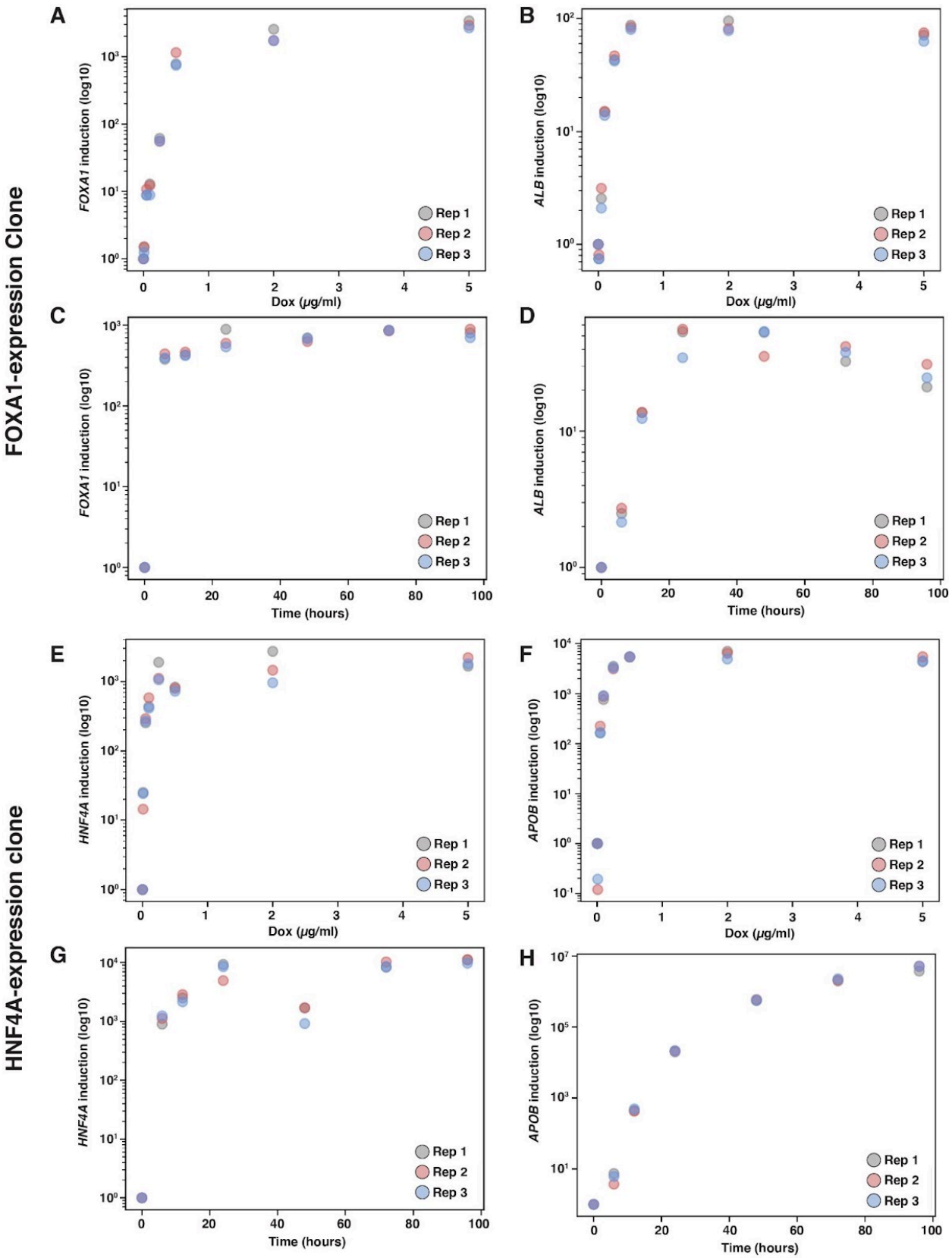
FOXA1-expression Clone
HNF4A-expression clone
A
B
C
D
E
F
G
H
FOXA1 induction (log10)
ALB induction (log10)
HNF4A induction (log10)
APOB induction (log10)
Dox (μg/ml)
Time (hours)
Rep 1
Rep 2
Rep 3

Figure 2.S1: Titration of doxycycline concentration and treatment time for TF and target gene induction. qPCR measurements made from RNA extracted from either the FOXA1 clonal line **(A-D)** or the HNF4A clonal line **(E-H)** that was treated with either increasing doxycycline concentrations or longer time periods. Expression is displayed as log10 fold induction over either 0 μg/ml doxycycline control (for concentration titration) or time 0 (for time titration). Each sample primer was normalized to the *HPRT* housekeeping gene. Doxycycline concentration titration measurements were made at 0, 0.01, 0.05, 0.1, 0.5, 2, and 5 μg/ml. Doxycycline treatment time measurements were made at 0, 6, 12, 24, 48, 72, and 96 hours.

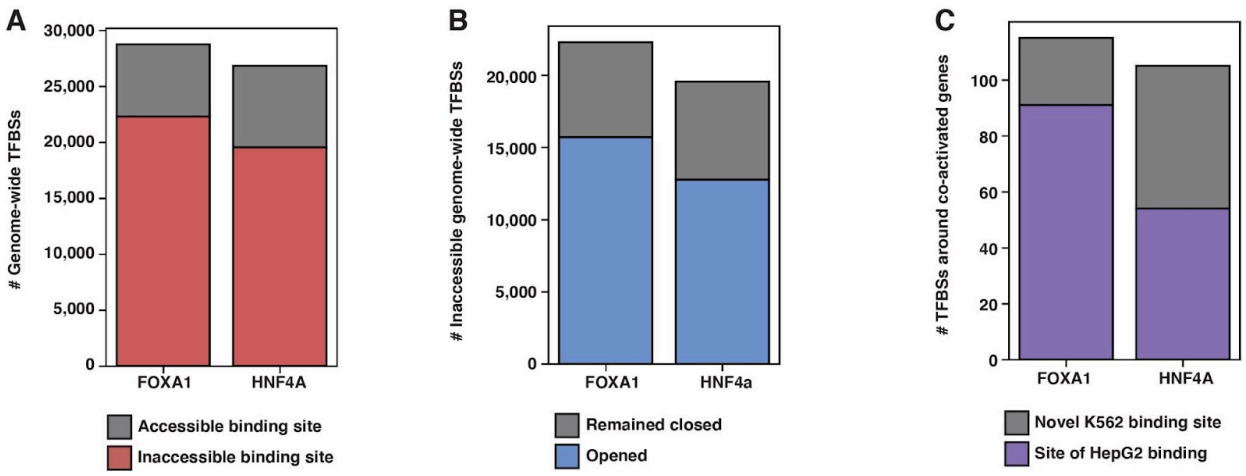

Figure 2.S2: Characterization of FOXA1 and HNF4A binding patterns in FOXA1-HNF4A clone. (A) The number of genome-wide FOXA1 or HNF4A transcription factor binding sites (TFBS) in the induced (+dox) cells that overlap with an ATAC-seq peak in the uninduced (-dox) cells ("Accessible binding site") or that do not overlap with an ATAC-seq peak in the uninduced (-dox) cells ("Inaccessible binding site"). **(B)** The number of inaccessible binding sites from (A) that overlap with an ATAC-seq peak in the induced (+dox) cells ("Opened") or that do not overlap with an ATAC-seq peak ("Remained closed"). **(C)** The number of FOXA1 or HNF4A binding sites within 50 kb of each FOXA1-HNF4A co-activated gene characterized as either a "HepG2 binding site," where the TFBS overlaps a TFBS of FOXA1 or HNF4A in HepG2 liver cells, or as a "Novel K562 binding site," where the TFBS does not overlap with a HepG2 binding site.

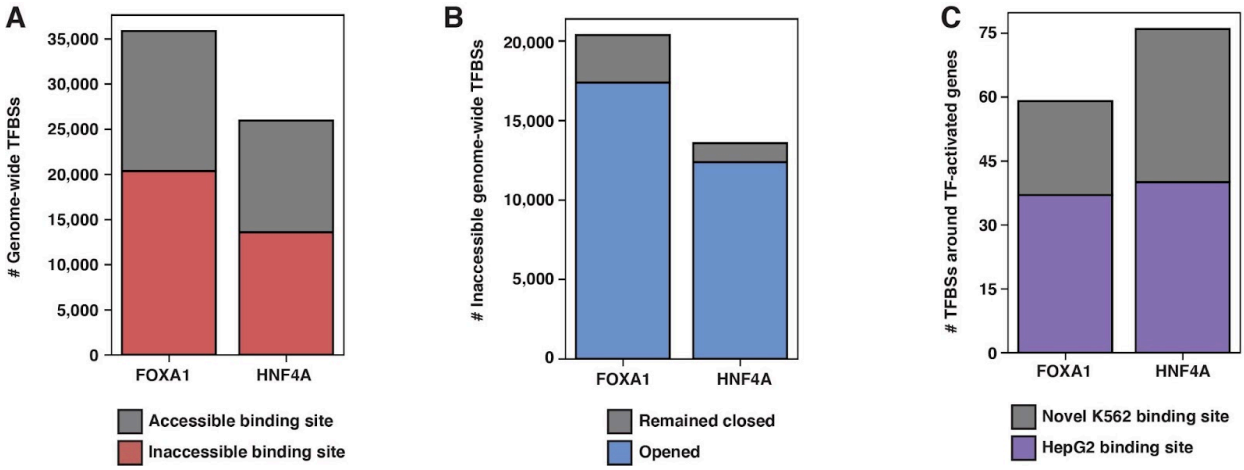

Figure 2.S3: Characterization of FOXA1 and HNF4A binding patterns in FOXA1 or HNF4A individual clones. (A) The number of genome-wide FOXA1 or HNF4A transcription factor binding sites (TFBS) in the induced (+dox) cells that overlap with an ATAC-seq peak in the uninduced (-dox) cells ("Accessible binding site") or that do not overlap with an ATAC-seq peak in the uninduced (-dox) cells ("Inaccessible binding site"). **(B)** The number of inaccessible binding sites from (A) that overlap with an ATAC-seq peak in the induced (+dox) cells ("Opened") or that do not overlap with an ATAC-seq peak ("Remained closed"). **(C)** The number of FOXA1 or HNF4A binding sites within 50 kb of each FOXA1- or HNF4A-activated gene characterized as either a "HepG2 binding site," where the TFBS overlaps a TFBS of FOXA1 or HNF4A in HepG2 liver cells, or as a "Novel K562 binding site," where the TFBS does not overlap with a HepG2 binding site.

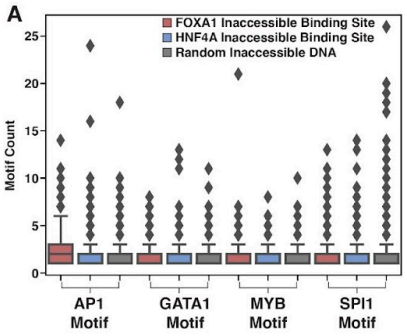

Figure 2.S4: K562 TF motif content in binding sites. (A) FIMO scans at p-value threshold 1e-3 for four most common proposed K562 PFs in either FoxA1 inaccessible binding sites (red), Hnf4a inaccessible binding sites (blue), or random equally-lengthed binding sites (gray).

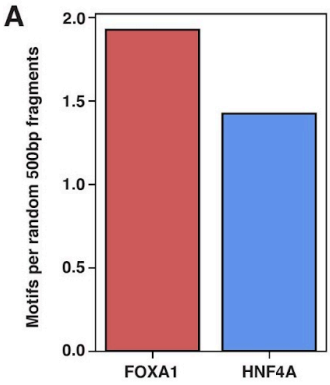

Figure 2.S5: FOXA1 and HNF4A motif scanning. (A) 1,000 random 200 bp fragments were generated using Bedtools and then scanned for FOXA1 and HNF4A motifs with FIMO using 1e-3 a p-value threshold. Total motif count was divided by the number of non-N containing random sequences (924) to identify motifs per random 200bp fragment.

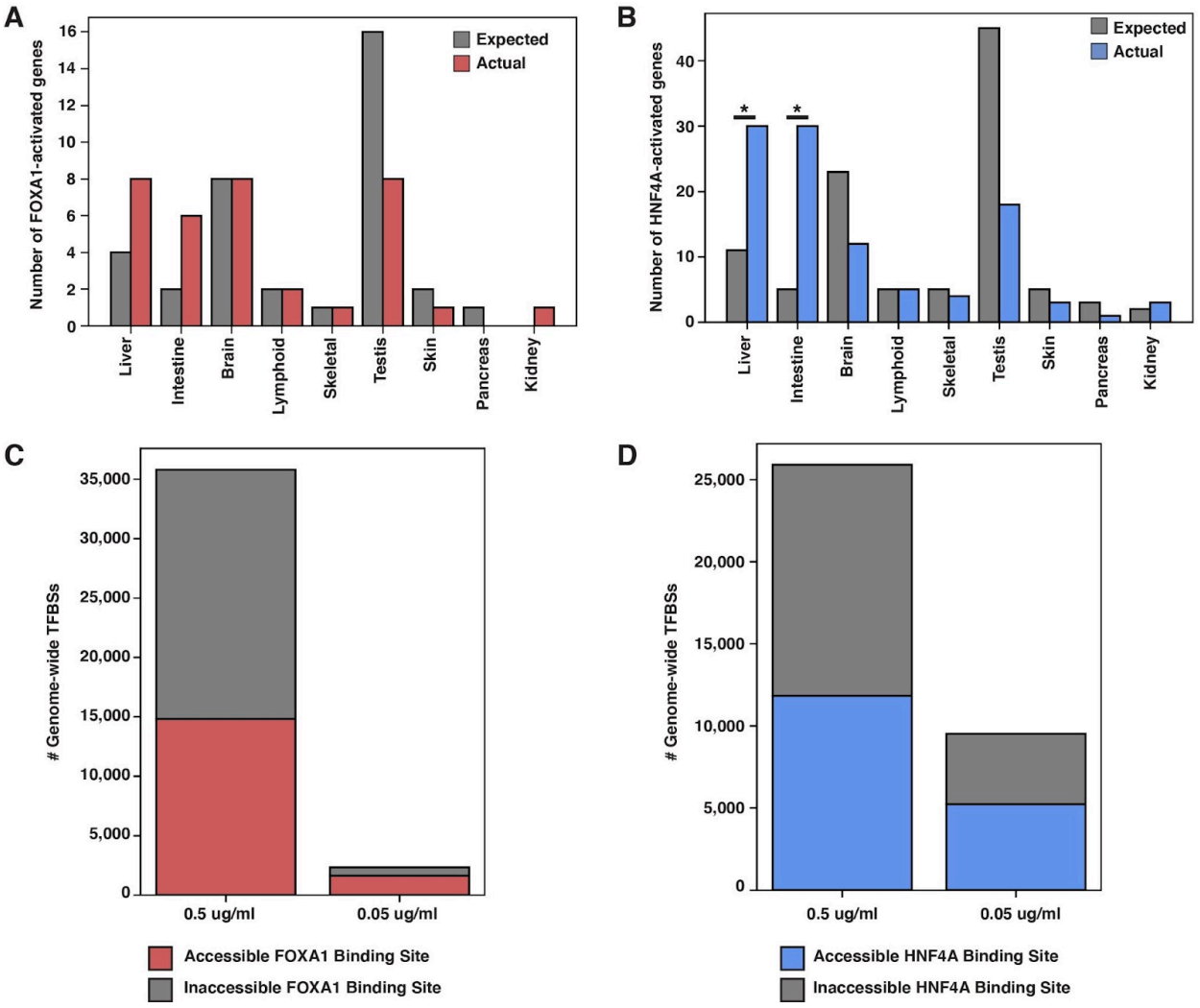

Figure 2.S6: Expression and binding at lower doxycycline induction. (A) The number of tissue-specific genes predicted from the hypergeometric distribution to be activated by FOXA1 at a lower doxycycline concentration (0.05 µg/ml) compared to the number actually activated. There are 242 total liver-enriched genes. **(B)** The number of tissue-specific genes predicted from the hypergeometric distribution to be activated by HNF4A at a lower doxycycline concentration (0.05 µg/ml) compared to the number actually activated. Liver- ($P < 10^{-5}$) and intestine-enrichment ($P < 10^{-14}$) are significant. There are 242 total liver-enriched genes and 122 total intestine-enriched genes. **(C-D)** Genome-wide FOXA1 **(C)** or HNF4A **(D)** binding sites classified as either events that occurred at sites that were accessible or inaccessible in the uninduced (-dox) state at 0.5 and 0.05 µg/ml doxycycline induction.

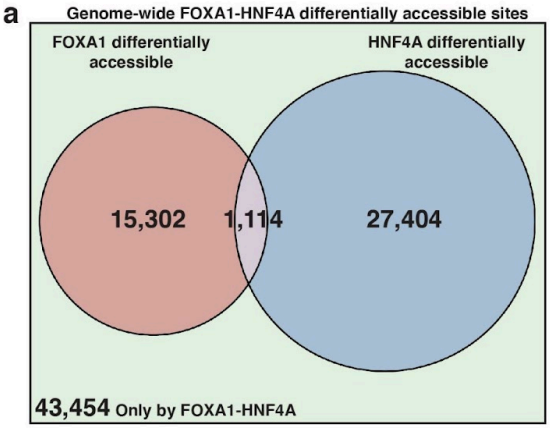

Figure 2.S7: Characterization of FOXA1-HNF4A differential accessibility. (A) Venn diagram of all FOXA1-HNF4A induced differentially accessible peaks categorized by whether the peak was also induced in the FOXA1 clone, HNF4A clone, neither, or both.

Supplementary Tables

Table 2.S1: FOXA1 Gene Ontology Analysis.

FOXA1 overrepresented biological process	Representative genes	FDR
Fibrinolysis	*SERPINF2, SERPING1, FGB, FGG, F2*	9.48e-07
Negative regulation of complement activation	*SERPINF2, VTN, F2*	4.36e-02
Regulation of heterotypic cell-cell adhesion	*APOA1, FGB, FGG*	6.24e-03
Acute phase response	*ITIH4, F2, SERPINF2*	5.75e-04
Platelet degranulation	*ALB, FGB, FGG, SERPINF2*	4.00e-06

Table 2.S2: HNF4A Gene Ontology Analysis.

HNF4A overrepresented biological process	Representative genes	FDR
Negative regulation of cholesterol import	*APOC3, APOA2*	5.92e-03
Negative regulation of VLDL particle remodeling/clearance	*APOA1, APOC3, APOA2*	1.97e-04
Chylomicron assembly/remodeling	*APOC2, APOC3, APOA1, APOA2*	2.22e-05
Tyrosine catabolic process	*HGD, HPD*	1.61e-02
Phospholipid efflux	*APOC2, APOC3, APOA1, APOA2*	4.14e-05

Table 2.S3: ATAC-sequencing quality summary statistics.

TF	Treatment	Replicate	Run Type	# Reads	# Peaks	FRIP
FOXA1	-dox	1	2x75	44,743,030	95,990	0.314
FOXA1	-dox	2	2x75	52,055,353	95,990	0.282
FOXA1	+dox	1	2x75	48,185,142	109,307	0.243
FOXA1	+dox	2	2x75	54,274,296	109,307	0.233
HNF4A	-dox	1	2x75	54,010,114	101,013	0.361
HNF4A	-dox	2	2x75	66,516,188	101,013	0.319
HNF4A	+dox	1	2x75	89,510,323	137,042	0.326
HNF4A	+dox	2	2x75	79,799,450	137,042	0.329
FOXA1-HNF4A	-dox	1	2x75	93,547,030	43,269	0.252
FOXA1-HNF4A	-dox	2	2x75	116,103,754	43,269	0.277
FOXA1-HNF4A	+dox	1	2x75	110,957,971	66,716	0.34
FOXA1-HNF4A	+dox	2	2x75	101,240,257	66,716	0.341

Table 2.S4: CUT&Tag sequencing quality summary statistics.

TF	Treatment	Antibody	Replicate	Run Type	# Reads	# Peaks	FRIP
FOXA1	+dox	FOXA1	1	2x150	19,489,387	42,734	0.369
FOXA1	+dox	FOXA1	2	2x150	23,025,631	51,994	0.377
HNF4A	+dox	HNF4A	1	2x150	15,154,729	37,233	0.331
HNF4A	+dox	HNF4A	2	2x150	15,615,127	31,864	0.269
FOXA1-HNF4A	+dox	FOXA1	1	2x150	22,946,339	33,819	0.386
FOXA1-HNF4A	+dox	FOXA1	2	2x150	22,739,194	34,145	0.37
FOXA1-HNF4A	+dox	HNF4A	1	2x150	18,242,414	41,187	0.313
FOXA1-HNF4A	+dox	HNF4A	2	2x150	16,269,780	26,193	0.222
FOXA1	+0.05dox	FOXA1	1	2x150	46,090,492	6,039	0.095
FOXA1	+0.05dox	FOXA1	2	2x150	24,625,780	3,218	0.075
HNF4A	+0.05dox	HNF4A	1	2x150	43,839,229	14,301	0.109

HNF4A	+0.05dox	HNF4A	2	2x150	44,493,248	13,552	0.109

Table 2.S5: Oligonucleotide sequences.

Name	Sequence
foxa1-v5-step1-R	agagggttagggataggcttaccacttgtattcaaaactggtcg
hnf4a-v5-step1-R	agagggttagggataggcttaccagcaacttgcccaaagcggc
v5-step2-F	ctacgtagaatcgagaccgaggagagggttagggataggctt
foxa1-pinducer-F	ccagcctccgcggccccgaaatgttgggcaccgtgaag
foxa1-pinducer-R	tgggacgtcgtatgggtattctacgtagaatcgagaccg
hnf4a-pinducer-F	ccagcctccgcggccccgaaatgcgactctccaaaacc
hnf4a-pinducer-R	tgggacgtcgtatgggtattctacgtagaatcgagacc
foxa1-qpcr-F	catgagacaagcgactggaa
foxa1-qpcr-R	tattaaaggaggccggtgtc
alb-qpcr-F	ctgcctgcctgttgccaaagc
alb-qpcr-R	ggcaaggtccgccctgtcatc
hnf4a-qpcr-F	aatgacacgtccccatcaga
hnf4a-qpcr-R	ggagtacatgtggttcttcc
apob-qpcr-F	agaggacagagccttggtggat
apob-qpcr-R	ctggacaaggtcatactctgcc
hprt-qpcr-F	tgacactggcaaaacaatgca
hprt-qpcr-R	ggtccttttcaccagcaagct

Chapter 3 – A Quantitative Metric of Pioneer Activity Reveals That HNF4A Has Stronger In Vivo Pioneer Activity Than FOXA1

A quantitative metric of pioneer activity reveals that HNF4A has stronger in vivo pioneer activity than FOXA1

Jeffrey L Hansen[1,2,3], Barak A Cohen[1,2*]

Affiliations

[1] The Edison Family Center for Genome Sciences and Systems Biology, School of Medicine, Washington University in St. Louis, Saint Louis, MO, USA.

[2] Department of Genetics, School of Medicine, Washington University in St. Louis, Saint Louis, MO, USA.

[3] Medical Scientist Training Program, Washington University in St. Louis, Saint. Louis, MO, USA

[*]Correspondence to: cohen@wustl.edu

We recently submitted this chapter as a manuscript to Nature Biotechnology. It will likely be modified from this version as it goes through the review process. I hope that the main claims do not change. Barak Cohen and I designed and analyzed the experiments and wrote the paper. I conducted the experiments. We may add an author if there is significant work that needs to be done during the review process.

3.1 – Abstract

We and others have suggested that pioneer activity–a transcription factor's (TF's) ability to bind and open inaccessible loci–is not a qualitative trait limited to a select class of pioneer TFs. We hypothesize that most TFs display pioneering activity that depends on the TF concentration and the motif content at their target loci. Here we present a quantitative measure of pioneer activity that captures the relative difference in a TF's ability to bind accessible versus inaccessible DNA. The metric is based on experiments that use CUT&Tag to measure binding of doxycycline (dox) inducible TFs. For each location across the genome we determine a "dox_{50}," the concentration of dox required for a TF to reach half-maximal occupancy. We propose that the ratio of a TF's average dox_{50} between ATAC-seq labeled inaccessible and accessible binding sites, its Δdox_{50}, is a measure of its pioneer activity. We measured Δdox_{50}'s for the endodermal TFs FOXA1 and HNF4A and show that HNF4A has a smaller Δdox_{50} than FOXA1, suggesting that HNF4A has stronger pioneer activity than FOXA1. We further show that FOXA1 binding sites with more copies of its motif have a lower Δdox_{50}, suggesting that strong motif content may compensate for weak pioneer activity. Our results suggest that Δdox_{50}s, or other similar measures that assess the difference in TF affinity for inaccessible and accessible DNA, are reasonable measures of pioneer activity.

3.2 – Introduction

Activating silent genes requires transcription factors (TFs) to bind and open DNA when their motifs are occluded by nucleosomes. Activating silent genes is postulated to involve two qualitatively different classes of TFs, pioneer factors (PFs) and non-pioneer factors (nonPFs) (Lisa Ann Cirillo et al. 2002; Iwafuchi-Doi and Zaret 2014). According to this hypothesis PFs bind to nucleosome-occluded DNA and make it accessible to nonPFs, which then recruit the cofactors required to activate transcription. However, we recently showed that both a canonical PF, FOXA1, and a nonPF, HNF4A, can independently bind, open, and then activate nearby genes (Hansen, Loell, and Cohen 2022), and many TFs possess unique ways of binding and opening nucleosomal DNA (F. Zhu et al. 2018; Swinstead et al. 2016; Miller and Widom 2003; Soufi et al. 2015; Yu and Buck 2019). From these data we propose that most TFs have quantifiable pioneer activity that depends on their nuclear concentrations and the motif content at their target loci. Here we present a metric that quantifies the pioneer activity of TFs at loci across the genome.

3.3 – Results

Definition of "Δdox_{50}" parameter for pioneer activity

An appropriate measure of pioneer activity should capture the relative difference of TF binding between inaccessible and accessible sites in the genome. In principle, we could compare the dissociation constant (K_d) of a TF at inaccessible and accessible sites as a measure of pioneer activity, since the K_d is the concentration of TF required to reach half maximal binding. In

practice, computing a K_d inside cells is impractical because it requires measuring the absolute concentration of a TF in the nucleus, in its proper post-translationally modified state. We propose a related measure that uses doxycycline-inducible (dox-inducible) TFs to compute the dox_{50}, the dox concentration required to reach half-maximal binding inside cells. By inducing TF levels over a wide range of dox concentrations and measuring the resulting binding by CUT&Tag (Kaya-Okur et al. 2019), we determine a dox_{50} for every location in the genome in parallel. The ratio of the average dox_{50} at inaccessible versus accessible sites, its Δdox_{50}, is a quantitative measure of a TF's pioneering activity. The smaller a TF's Δdox_{50}, the less its binding is reduced at inaccessible DNA (Figure 3.1A). Because the measurements at inaccessible and accessible sites are made at the same time in the same nucleus, the dox concentrations (or TF concentrations) cancel out, allowing us to compare the Δdox_{50}s of different TFs to each other (Man and Stormo 2001). This strategy allows us to circumvent the challenge of measuring effective nuclear TF concentration while maintaining the physiological relevance of our in vivo pioneer activity measurements.

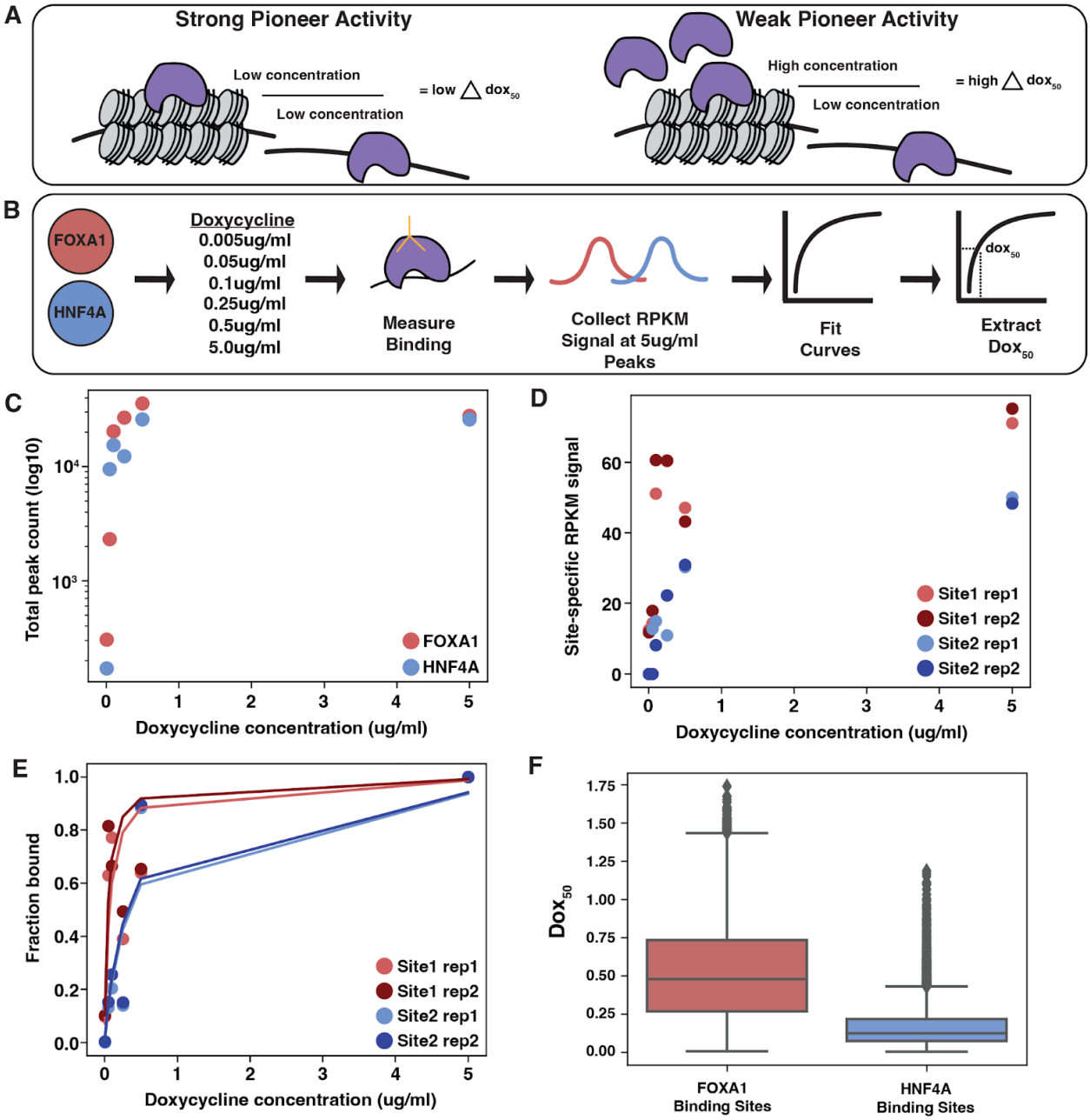

Figure 3.1. Experimental design to calculate TF dox$_{50}$ values across the genome. (A) Strong pioneer activity leads to smaller differences in the dox concentration needed to achieve binding at inaccessible versus accessible sites, and vice versa. **(B)** We induced FOXA1 or HNF4A across a 1,000-fold dox range, measured binding, collected binding signal at a reproducible set of binding sites and then extracted a dox$_{50}$ for each site. **(C)** Number of total peaks for each TF across each dox induction level. **(D)** Replicate RPKM binding signal at two example genomic sites. **(E)** Replicate fitted lines at two example genomic sites. **(F)** Full distribution of dox$_{50}$ values for each TF.

Measurement of dox$_{50}$ for FOXA1 and HNF4A

FOXA1 and HNF4A are liver TFs that are commonly used to reprogram embryonic fibroblasts to endoderm progenitor cells (Biddy et al. 2018; Sekiya and Suzuki 2011). FOXA1 is a canonical PF and HNF4A a nonPF, and the two are suggested to work in a collaborative and sequential fashion to activate their target genes (Horisawa et al. 2020; Lisa Ann Cirillo et al. 2002). We previously tested FOXA1 and HNF4A's behavior in an ectopic setting by expressing them within K562 blood cells, a lineage in which neither TF is expressed and that should present the TFs with unique complements of chromatin and cofactors. We created clonal K562 lines that expressed either inducible FOXA1 or HNF4A and showed that both TFs could independently bind and open inaccessible chromatin and activate nearby genes (Hansen, Loell, and Cohen 2022).

Based on the ability of FOXA1 and HNF4A to independently bind, open, and activate in an ectopic cell line, we expected both TFs would have similar pioneer activity. To test this prediction we attempted to measure each TF's Δdox_{50} using the same dox-inducible FOXA1 or HNF4A K562 lines as in our previous work (Hansen, Loell, and Cohen 2022) (Figure 3.1B). We first treated each TF line with a 1,000-fold range of dox (0.005, 0.05, 0.25, 0.1, 0.5, and 5.0µg/ml) and measured resultant binding. Read normalized binding signal (RPKM) was highly correlated between replicates (Figure 3.S1). FOXA1 and HNF4A appear to be expressed at reasonably similar levels at each dox level as each TF bound to similar numbers of sites in each condition (Figure 3.1C). We then collected the overlapping set of binding sites between each TF's replicates in the 5.0µg/ml sample and at each site plotted the read normalized signal (RPKM) from the other induction levels (Figure 3.1D). Generally the binding patterns follow

predicted saturation binding kinetics (Figure 3.S2). We then fit Equation 3.1 to these distributions.

$$\text{Fraction bound} = \frac{1}{1 + \frac{dox_{50}}{[dox]}} \qquad \text{(Equation 3.1)}$$

In order to fit Equation 3.1, we normalized each site's RPKM signal to the signal in the 5.0µg/ml sample to convert our measurements into fractional binding (Figure 3.1E). We found that at some sites the binding signal peaked prior to the 5.0µg/ml sample (fraction bound > 1 in any of the first five induction levels). We removed these sites to prevent poor fitting. This left us with 11,557 FOXA1 binding curves and 5,940 HNF4A binding curves with highly similar fitted lines across replicates (Figure 3.1F, Fig. 3.S3). We extracted dox_{50} values from these lines and found similar results between replicates (Figure 3.S4) and so we averaged each site's replicate dox_{50} value for the remaining analyses. The resulting distributions of each TF's genome-wide dox_{50} values show that FOXA1 has a much larger variance in dox_{50} values than HNF4A (Figure 3.1F), suggesting that FOXA1 binding generally depends more on the genomic environment than HNF4A.

Measurement of Δdox_{50} for FOXA1 and HNF4A

The dox_{50} distributions in Figure 3.1 suggest that HNF4A may bind more consistently across the genome but do not explicitly measure their pioneer activities, the difference in each TFs ability to bind at inaccessible versus accessible sites. We therefore classified each site as either inaccessible or accessible based on ATAC-seq (Buenrostro et al. 2015) peaks collected in these

cell lines before induction. Of FOXA1's 11,557 peaks, 1,930 were in accessible regions and 9,627 were in inaccessible regions (10,120 accessible before filtering, 17,644 inaccessible before filtering). Of HNF4A's 5,940 peaks, 2,135 were in accessible regions and 3,805 were in inaccessible regions (16,137 accessible before filtering, 16,507 after filtering). Comparing the dox_{50} distributions between inaccessible and accessible sites revealed that the binding of HNF4A is less affected by inaccessible DNA than FOXA1 (Figure 3.2A). We then computed a Δdox_{50} for each TF by dividing the average dox_{50} for inaccessible sites by the average dox_{50} for accessible sites. This analysis showed that HNF4A has a lower Δdox_{50} than FOXA1 (Table 3.1). This result holds when we include those sites that were filtered out because they peaked at lower concentrations (Table 3.1, Fig. S5).

We next considered whether the motif content at each binding site affected the Δdox_{50}. We showed previously that both TF concentration and motif content affect FOXA1 and HNF4A's pioneer activity and speculated that any parameter that affects occupancy will be important (Hansen, Loell, and Cohen 2022). Therefore, we further subset our binding sites into those that had less than 2, between 2-4, or more than 4 motifs and re-plotted the dox_{50} distributions and re-calculated Δdox_{50}s. Higher motif content allowed FOXA1 to bind more consistently between inaccessible and accessible sites and thus lowered FOXA1's Δdox_{50} (Figure 3.2D, Table 3.1). In contrast, motif count had little effect on HNF4A (Figure 3.2D, Table 3.1). We conclude that HNF4A has stronger pioneer activity in K562 cells and that weaker pioneer activity can be compensated by strong motif content.

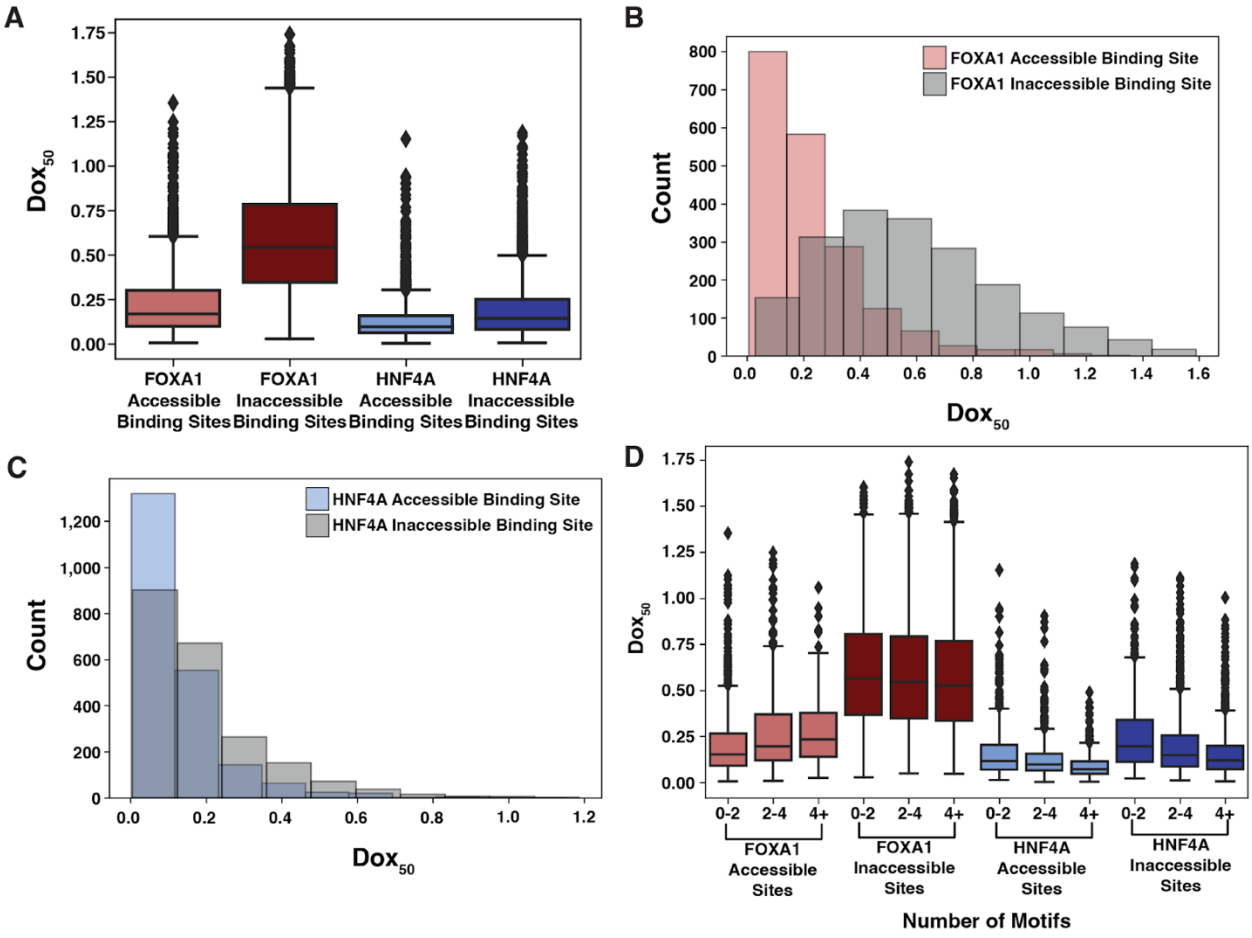

Figure 3.2. HNF4A has a smaller Δdox$_{50}$ than FOXA1. (A) Distributions of dox$_{50}$ estimates extracted from binding curves at FOXA1 accessible binding sites (n = 1,930), FOXA1 inaccessible binding sites (n = 9,627), HNF4A accessible binding sites (n = 2,135), and HNF4A inaccessible binding sites (n = 3,805). **(B-C)** Distributions from FOXA1 **(B)** and HNF4A **(C)** shown in histogram form. **(D)** Same plot as (A) but each genomic binding site is binned by whether the site has < 2, >= 2 but < 4, or >= 4 motifs as called by FIMO (p = 1e-3).

Table 3.1. Δdox$_{50}$s for FOXA1 and HNF4A across different types of binding sites.

TF	All sites	Filtered sites	<2 motifs	2-4 motifs	>4 motifs
FOXA1	5.23	2.56	2.96	2.16	2.02
HNF4A	1.81	1.47	1.55	1.56	1.70

Chromatin modifications explain some dox$_{50}$ variance

We built a linear model (Equation 3.2) to try to explain the variance in dox$_{50}$s for FOXA1 and HNF4A where C(Accessibility) is each binding site's accessibility prior to TF induction.

Accessibility explained 17% of the variance in FOXA1's dox_{50}s but only 4% of HNF4A's. While these data further underscore the greater role that accessibility plays on FOXA1 binding than HNF4A, they also reveal that most of the variance in dox_{50} values between genomic loci must be explained by some other variable.

$$\mathbf{Dox_{50} \sim C(Accessibility)} \qquad \textbf{(Equation 3.2)}$$

We hypothesized that some of the remaining variance may be explained by the different chromatin modifications present at different target loci and predicted that binding sites with active marks would have lower dox_{50} distributions (easier binding) and binding sites with silent marks would have higher dox_{50} distributions (harder binding). We further subset each TF's accessible or inaccessible binding sites into those that overlap common K562 marks (Zhang et al. 2020). H3K4me1 marks enhancers (Heintzman et al. 2007), H3K27Ac marks activity (Creyghton et al. 2010), and H3K9me3 and H3K27me3 are two modifications shown previously to suppress pioneer activity (Mayran et al. 2018). The accessible sites overlapped much more often with active marks than silencing marks, and vice versa, and we found that no FOXA1 or HNF4A accessible sites were marked with H3K9me3 (Figure 3.3).

As we predicted, FOXA1 or HNF4A binding sites that overlapped H3K27Ac or H3K4me1 chromatin modifications had lower dox_{50} distributions than those that overlapped H3K9me3 or H3K27me3 (Figure 3.3). These effects were present even after we subset binding sites by accessibility, suggesting that the chromatin modifications can affect binding in ways that are separable from the effects of accessibility. However, when we individually added each chromatin

modification (plus an interaction term) to the model in Equation 3.2, we found that accounting for these marks did not have large effects on the ability of the model to predict dox_{50} values for either TF. H3K27ac levels explained 2% of FOXA1's dox_{50} variance, H3K4me1 explained 1%, and H3K27me3 explained <1%. For HNF4A, H3K27ac explained 2%, H3K4me1 explained 2%, and H3K27me3 explained <1%. All interaction terms were negligible. Together these data suggest that something besides the epigenetic landscape of loci is having a large effect on the pioneering activity of TFs.

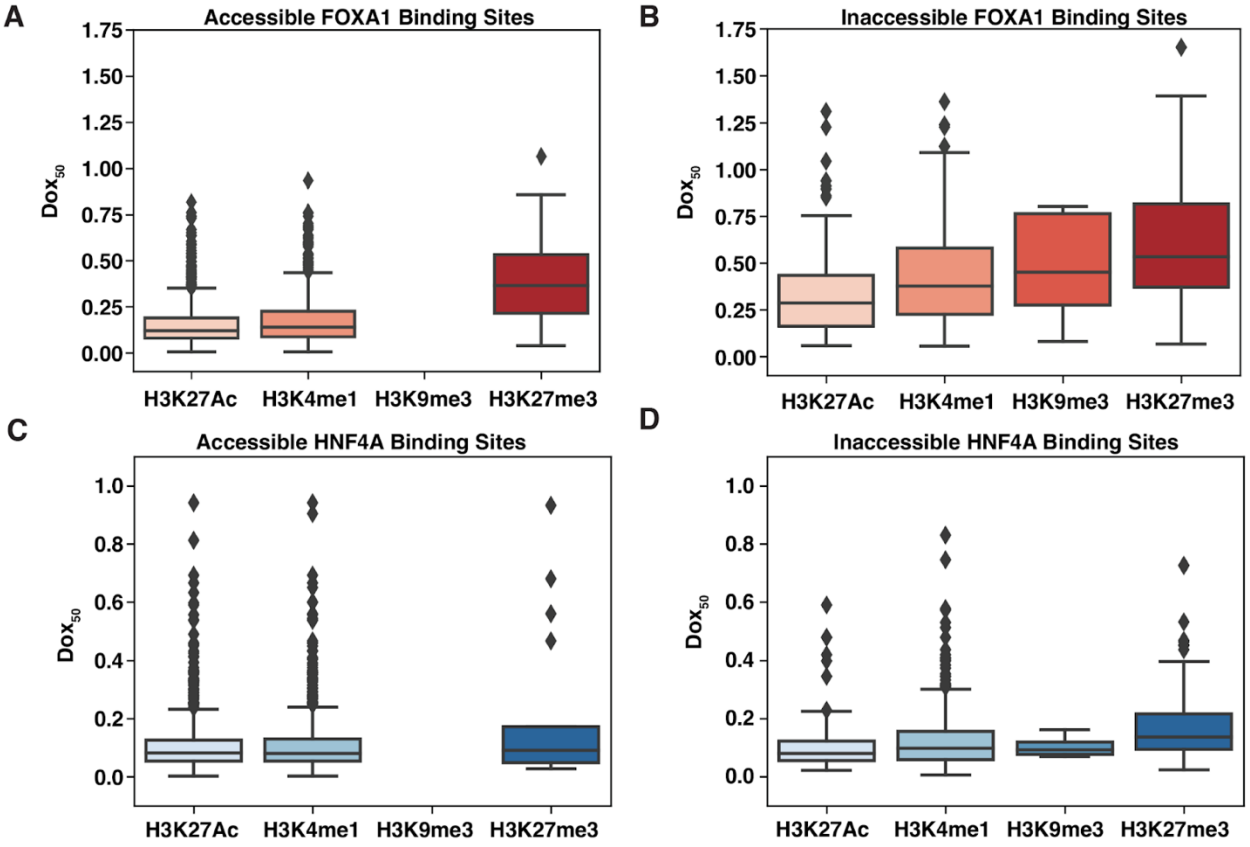

Figure 3.3. Dox_{50} distributions across different chromatin modifications. (A) Dox_{50} values for FOXA1 accessible binding sites that overlapped H3K27AC (n = 1,288), H3K4me1 (n = 755), H3K9me3 (n = 0), and H3K27me3 (n = 21). **(B)** Dox_{50} values for FOXA1 inaccessible binding sites that overlapped H3K27AC (n = 203), H3K4me1 (n = 352), H3K9me3 (n = 12), and H3K27me3 (n = 277). **(C)** Dox_{50} values for HNF4A accessible binding sites that overlapped H3K27AC (n = 1,147), H3K4me1 (n = 1,111), H3K9me3 (n = 0), and H3K27me3 (n = 17). **(D)** Dox_{50} values for HNF4A inaccessible binding sites that overlapped H3K27AC (n = 140), H3K4me1 (n = 416), H3K9me3 (n = 4), and H3K27me3 (n = 135).

FOXA1 behaves anti-cooperatively at a subset of accessible binding sites

While examining individual binding sites and their fitted curves, we observed a repeating pattern at a subset of genomic locations where the binding signal increased to a peak at the third (0.1µg/ml) or fourth (0.25µg/ml) induction level and then decreased at the highest dox concentration, suggesting anti-cooperative behavior (Figure 3.4A, Figure 3.S6). To quantify the prevalence of anti-cooperative binding, we sampled 10,000 peaks from FOXA1 or HNF4A inaccessible or accessible binding sites and then counted how many displayed saturation behavior (peak at 5µg/ml, Figure 3.S2) and how many displayed anti-cooperative behavior (peak at 0.1µg/ml or 0.25µg/ml, Figure 3.S6). We found that the anti-cooperative behavior occurs most often at accessible FOXA1 binding sites (Figure 3.4B-C). Anti-cooperative behavior does not appear to depend on the number of motifs at each peak (Figure 3.4D) or the length of each peak (Figure 3.4E).

We considered whether another TF might be contributing to anti-cooperative behavior by searching for enriched motifs in either saturation-type accessible FOXA1 binding sites or anti cooperative-type sites. While FOXA1 sites were enriched in both types of loci (Figure 3.4F), the AP1 motif was only enriched at anti-cooperative sites (Figure 3.4G). AP1 is an important K562 TF that exhibits some pioneer activity (Biddie et al. 2011). The results suggest that a genetic interaction between FOXA1 and AP1 underlies anti-cooperative behavior at accessible loci.

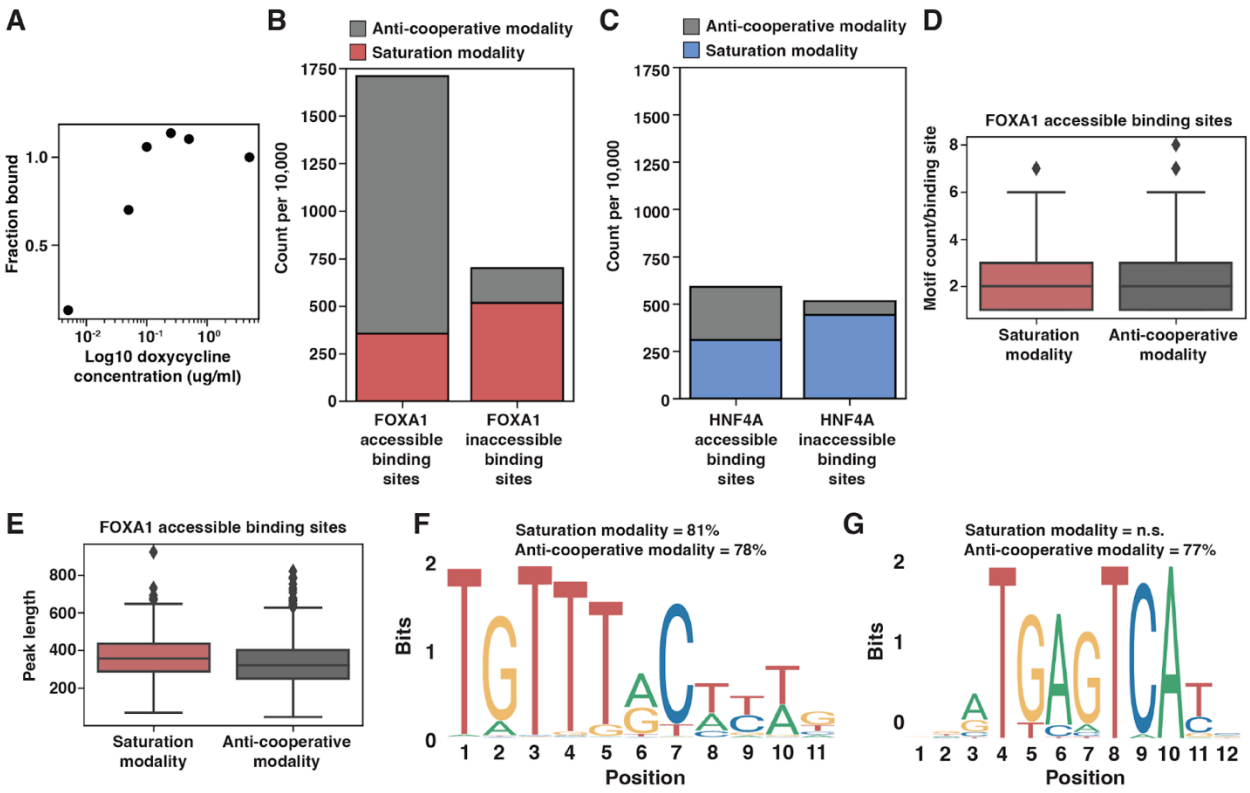

Figure 3.4. Characterization of anti-cooperative binding behavior. (A) Example binding curve at a single genomic site that exhibits anti-cooperative behavior. **(B-C)** A sample of 10,000 FOXA1 **(B)** or HNF4A **(C)** accessible (left bar) or inaccessible (right bar) binding sites colored by if they display saturation binding behavior (gray) or anti-cooperative binding behavior (red, see). **(D)** FOXA1 motif count between the accessible binding sites from (B) that display either saturation or anti-cooperative binding behavior. Motifs were called from FIMO with a p-value threshold of 1e-3. **(E)** Binding peak length between the accessible binding sites from (B) that display either saturation or anti-cooperative binding behavior. **(F)** The most enriched motif discovered in FOXA1 accessible saturation and anti-cooperative peaks was FOXA1 (JASPAR MA0148.1). It is significantly enriched for both the saturation behavior (p = 2.19e-001) and anti-cooperative behavior (p = 1.23e-01). **(G)** The second most enriched motif discovered in FOXA1 accessible anti-cooperative peaks was AP1 (JASPAR MA1141.1). It was not discovered in the saturation behavior peaks. It is significantly enriched for only the anti-cooperative behavior (p = 1e-008).

3.4 – Discussion

Given a definition of pioneer activity that is a TF's ability to bind at inaccessible genomic locations, we suggest that the Δdox_{50} is a quantitative measure of this activity. We measured the Δdox_{50} of FOXA1 and HNF4A in K562 cells and showed that HNF4A has stronger pioneer activity in this cell type than FOXA1. However, both TFs showed a range of dox_{50} values across

the genome, which demonstrates that a TFs pioneer activity may depend on accessibility, native chromatin marks, and other factors. Some of these differences are explained by the motif content at different locations, suggesting that low pioneer activity can be overcome by strong motif content. While our work shows that the pioneering activity of a TF can vary across the genome, what accounts for this variation across sites remains mostly unexplained. DNA accessibility had the largest effect on pioneer activity but only explained 17% of the variance in dox_{50} values. We speculate that much of the remaining variance in dox_{50} values might be explained by interactions with other specific TFs or with the general transcription machinery that can differ across the genome.

Our work supports the hypothesis that pioneer activity is not a qualitative trait limited to a few TFs, but rather a quantitative property of TFs that manifests differently depending on the TF and the genomic environment (Garcia et al. 2019; F. Zhu et al. 2018; Hansen, Loell, and Cohen 2022; Lisa Ann Cirillo et al. 2002; Soufi et al. 2015; Yu and Buck 2019). Pioneer activity as a quantitative trait fits with data showing that TFs can gain pioneer activity when expressed at high levels or when cells are forced into replication (Yan, Chen, and Bai 2018) and that TFs can lose pioneer activity when expressed at lower levels (Hansen, Loell, and Cohen 2022).

In vitro, FOXA1 has higher affinity (a lower K_d) for naked DNA than HNF4A (Jiang, Lee, and Sladek 1997; Garcia et al. 2019; Rufibach et al. 2006), and yet HNF4A has stronger pioneer activity (a lower Δdox_{50}) than FOXA1 in K562 cells. These results demonstrate that pioneer activity is not solely a function of the affinity of a TF's DNA-binding domain for its cognate motif. Inside cells, pioneer activity likely depends on the interactions a TF makes with other TFs

and with cofactors. Because these interactions will differ in different cell types, a TF's pioneering activity is also likely to depend on the cell type in which it is expressed and the co-bound TFs present at certain locations.

At some locations in the genome, an interaction between FOXA1 and AP1 appears to have a dramatic effect on FOXA1 activity. In the presence of AP1 sites, FOXA1 displays anti-cooperative binding dynamics where occupancy decreases at the highest levels of FOXA1 expression. We speculate that at these sites monomers of FOXA1 interact with AP1 to potentiate binding, whereas dimers of FOXA1 cannot cobind with AP1. In this model, high concentrations of FOXA1 favor its dimeric form which accounts for the loss of binding at these sites when FOXA1 is expressed at high levels. Regardless of the mechanism underlying anti-cooperative behavior, our results show that pioneer activity can be modified by the interactions a TF makes inside cells. Thus, pioneer activity is contingent on many properties of a TF including its levels, its intrinsic affinity for its motif, the motif content at its targets, and the different interactions it makes with other proteins when bound at different locations. Given these contingencies we suggest that most TFs will display some degree of pioneer activity and that the Δdox_{50}, or a related metric, will be a useful metric to quantify it.

3.5 – Materials and Methods

Cell lines

We grew K562 cells (ATCC CCL-243, Manassas, VA) in Iscove's Modified Dulbecco Serum supplemented with 10% fetal bovine serum, 1% penicillin-streptomycin and 1% non-essential

amino acids. We used these cell types to generate clonal FOXA1 and HNF4A lines, as described below. For each of our functional assays, we split each line into replicate flasks, treated with doxycycline (dox) (Sigma #D9891-1G), and then waited 24 hours to extract RNA or nuclei. We used doses of 0.005µg/ml, 0.05µg/ml, 0.1µg/ml, 0.25µg/ml, 0.5µg/ml, and 5µg/ml for our dox_{50} experiments.

Cloning, production, and infection of viral vectors

We used FOXA1 and HNF4A K562 clonal lines and lentiviral vectors carrying inducible FOXA1 and HNF4A ORFs as described previously (Hansen, Loell, and Cohen 2022).

Sequencing library preparations and analysis

We prepared sequencing libraries and analyzed the two replicates of CUT&Tag as described previously (Hansen et al. 2022). In our previous work we already used ATAC-seq to measure the uninduced (-dox) accessibility in the FOXA1 and HNF4A K562 lines (Hansen et al. 2022). Because we used the same clones to perform these experiments, we re-used these data as uninduced accessibility. We also had already sequenced CUT&Tag libraries for the 0.5ug/ml and 0.05ug/ml doxycycline induction levels and re-used these data as well.

Binding curve analysis

We first established a set of all possible binding sites for each TF by creating a list of binding sites in the sample with the highest dox induction concentration (5µg/ml). We subset this list into those accessible binding sites (called accessible peak in the -dox uninduced condition) and inaccessible binding sites (absence of called accessible peak). Then we used the

multiBigwigSummary from the deepTools suite (Ramírez et al. 2016) to count the normalized read intensity at each peak from each induction level. We normalized each induction level to the read intensity at the highest induction level in order to convert read intensity into fraction bound.

With these data, we fit a binding curve using SciPy curvefit (Virtanen et al. 2020) to the equation (Equation 3.1) where dox_{50} is unknown and represents a binding affinity parameter similar to K_d and where [dox] is the concentration of dox used to induce TF expression. When we plotted examples of randomly selected genomic sites and examined the binding curves, we noticed that at some sites, binding peaked (fraction bound >= 1) prior to the highest concentration. In these cases, the fit line estimated a negative dox_{50}. For this reason, we filtered out any site that peaked prior to the sample with the highest dox concentration. We also estimated dox_{50} distributions without this filtering step and found similar distributions. (Figure 3.S2).

In order to quantify the early peak, or "anti-cooperative" behavior that we observed, we classified a binding site as exhibiting a "saturation binding" modality if only the highest dox concentration had a fraction bound of 1, and then each subsequent lower concentration had a lower fraction bound. We classified a binding site as exhibiting an "anti-cooperative" modality if the site peaked at either the third (0.1µg/ml) or fourth (0.25µg/ml) dox concentrations and then declined in each direction.

We calculated reproducibility in three ways. We first showed that the binding signal was reproducible by plotting the RPKM signal from each replicate for each of the concentrations at all of the binding sites collected as described above. We then showed that the lines fit similarly

between replicates by both replicates' binding signal and fit binding curves at many different randomly chosen genomic sites and showing that the lines look similar. And finally we showed that the distributions of dox_{50}s from each replicate were highly overlapping. After showing these, we averaged the dox_{50} from each replicate at each site and used the average value moving forward.

Motif analysis

To discover or count motifs in binding sites, we extracted the sequence from each CUT&Tag binding peak and then used XSTREME (Grant and Bailey 2021) for de novo motif discovery and FIMO (Grant, Bailey, and Noble 2011) for specific motif occurrence counting. We used 1e-3 as a p-value threshold and JASPAR (Fornes et al. 2020) PWMs for FOXA1 (MA0148.1), HNF4A (MA0114.2), and AP-1 (MA1141.1). We used these motif counts to subset the FOXA1/HNF4A accessible/inaccessible peaks into those with less than 2 motifs, more than 2 but less than 4, or 4 or more, and then re-ran the analysis (Figure 3.S4).

Chromatin modifications analysis and modeling

We used previously published datasets of histone ChIP-seq (J. Zhang et al. 2020) to identify patterns of H3K27Ac, H3K4me1, H3K9me3, and H3K27me3 marks. We used BEDTools (Quinlan and Hall 2010) to overlap FOXA1 or HNF4A's binding sites with these marks. We then used python's statsmodels to run ANOVA analyses on ordinary least squares linear regressions. Each reported variance is the parameter's sum of squares contribution divided by the total sum of squares.

Data Availability

All genomic sequencing data have been deposited on Gene Expression Omnibus (GEO) under accession number GSE204726.

3.6 – Acknowledgements

We thank members of the Cohen Lab for reading and critiquing the manuscript and for helpful discussion; Jessica Hoisington-Lopez and MariaLynn Crosby in the DNA Sequencing Innovation Lab for assistance with high-throughput sequencing; the Genome Engineering and iPSC Center for allowing us to use their Sony Flow Cytometer for cell sorting; and Mingjie Li in the Hope Center Viral Vectors Core for assistance with producing lentiviral expression vectors. This work was supported by grants from the National Institutes of Health: R01GM092910 (Dr. Barak Cohen), T32HG000045 (Dr. Michael Brent, Washington University in St. Louis Genome Analysis Training Program), and T32GM007200 (Dr. Wayne Yokoyama, Washington University in St. Louis Medical Scientist Training Program).

Author Contributions

J.L.H. and B.A.C. designed the overall project. J.L.H. conducted experiments. J.L.H conducted analysis. J.L.H. and B.A.C. wrote the manuscript.

Competing Interests

The authors declare no competing interests.

3.7 – Supplementary Information

Supplementary Figures

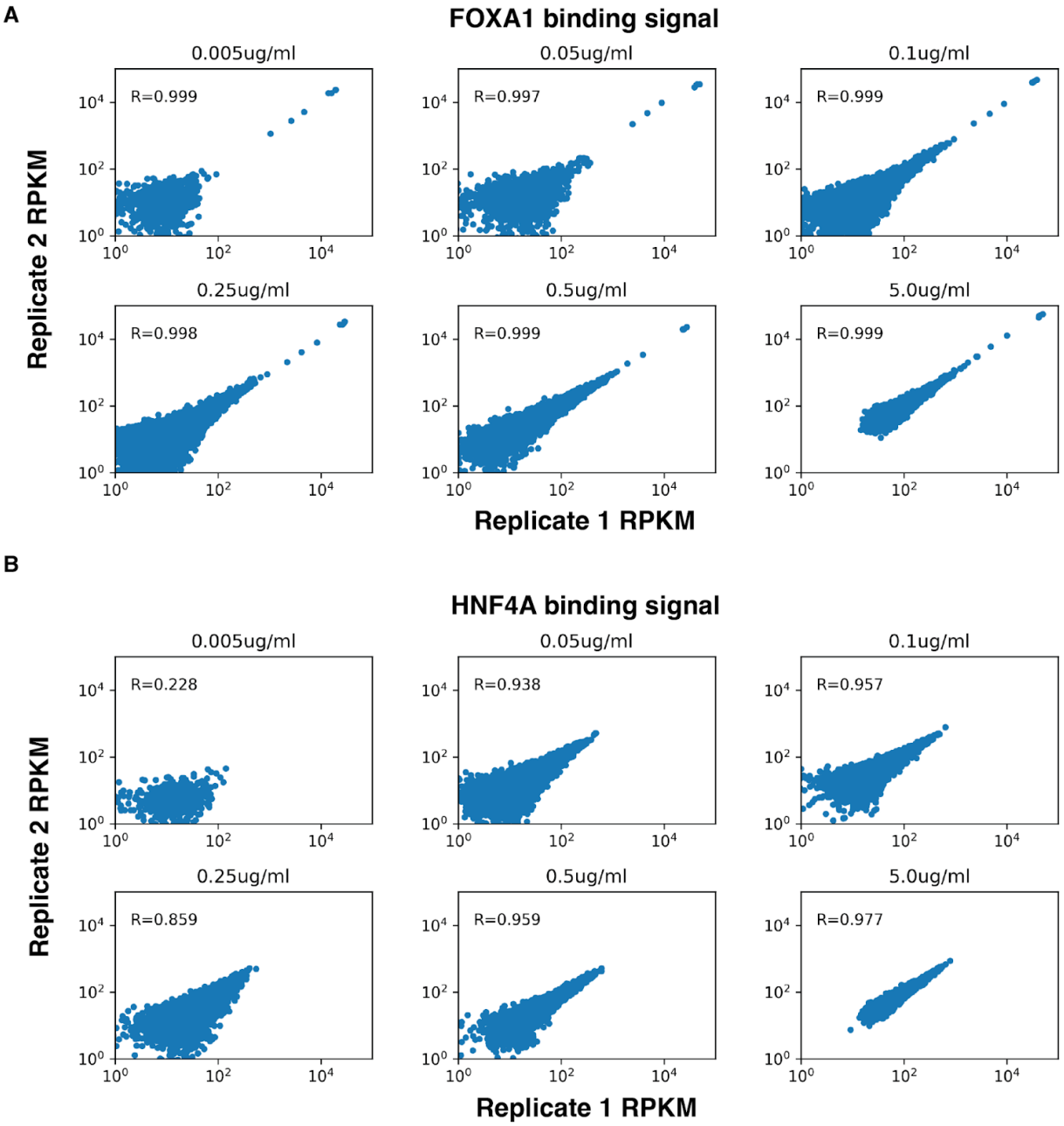

Figure 3.S1. Reproducibility of binding signal. RPKM signal from each replicate of CUT&Tag data across each TF across each dox induction concentration. Pearson's R correlation displayed on each graph.

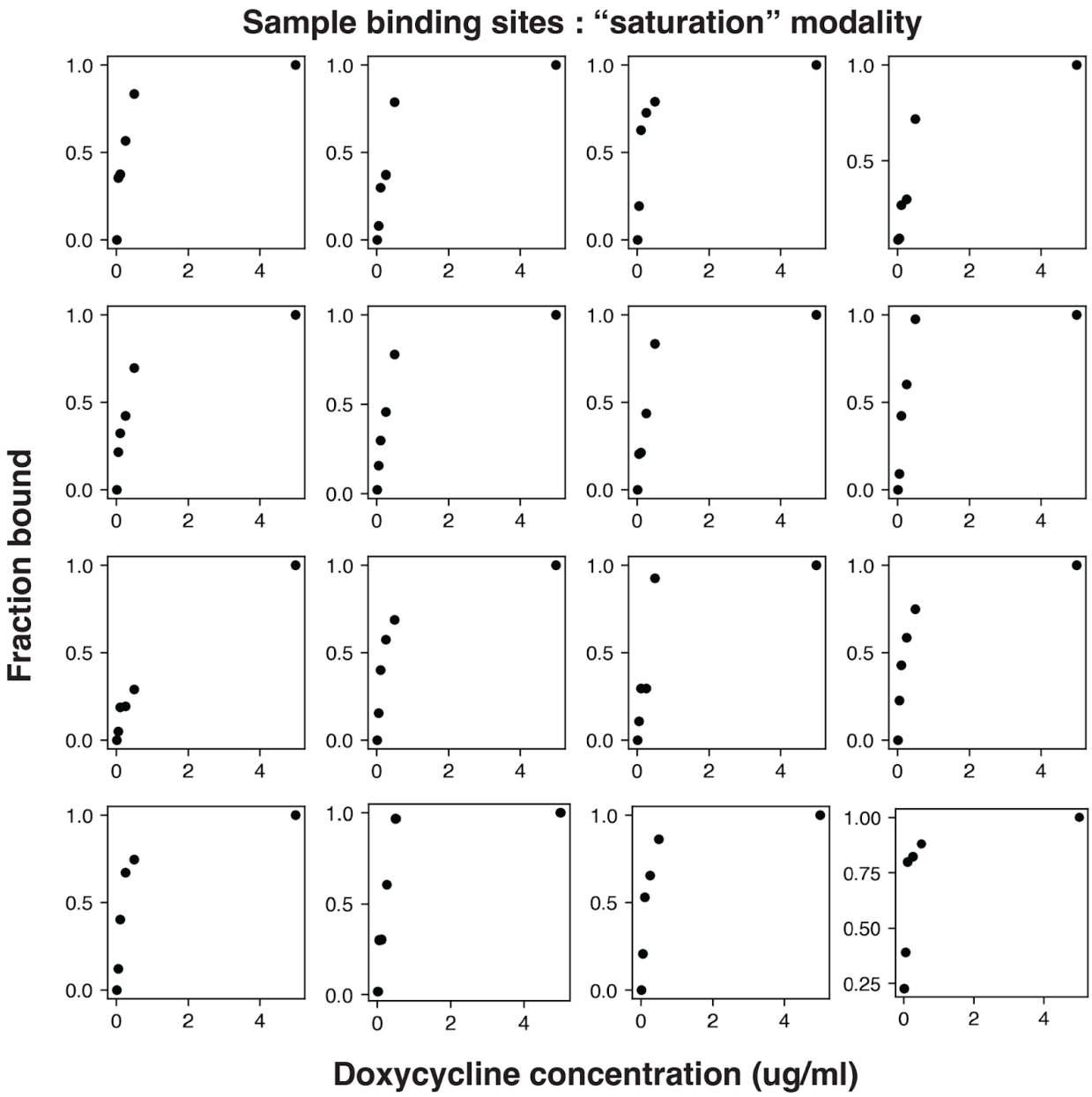

Figure 3.S2. Common saturation behavior binding pattern. 16 examples from different genomic sites showing saturating binding signal as dox induction increases. Signal is first read normalized (RPKM) and then normalized to the signal at the highest concentration. These sites were sampled from FOXA1 accessible binding sites but are common across inaccessible and HNF4A binding sites as well.

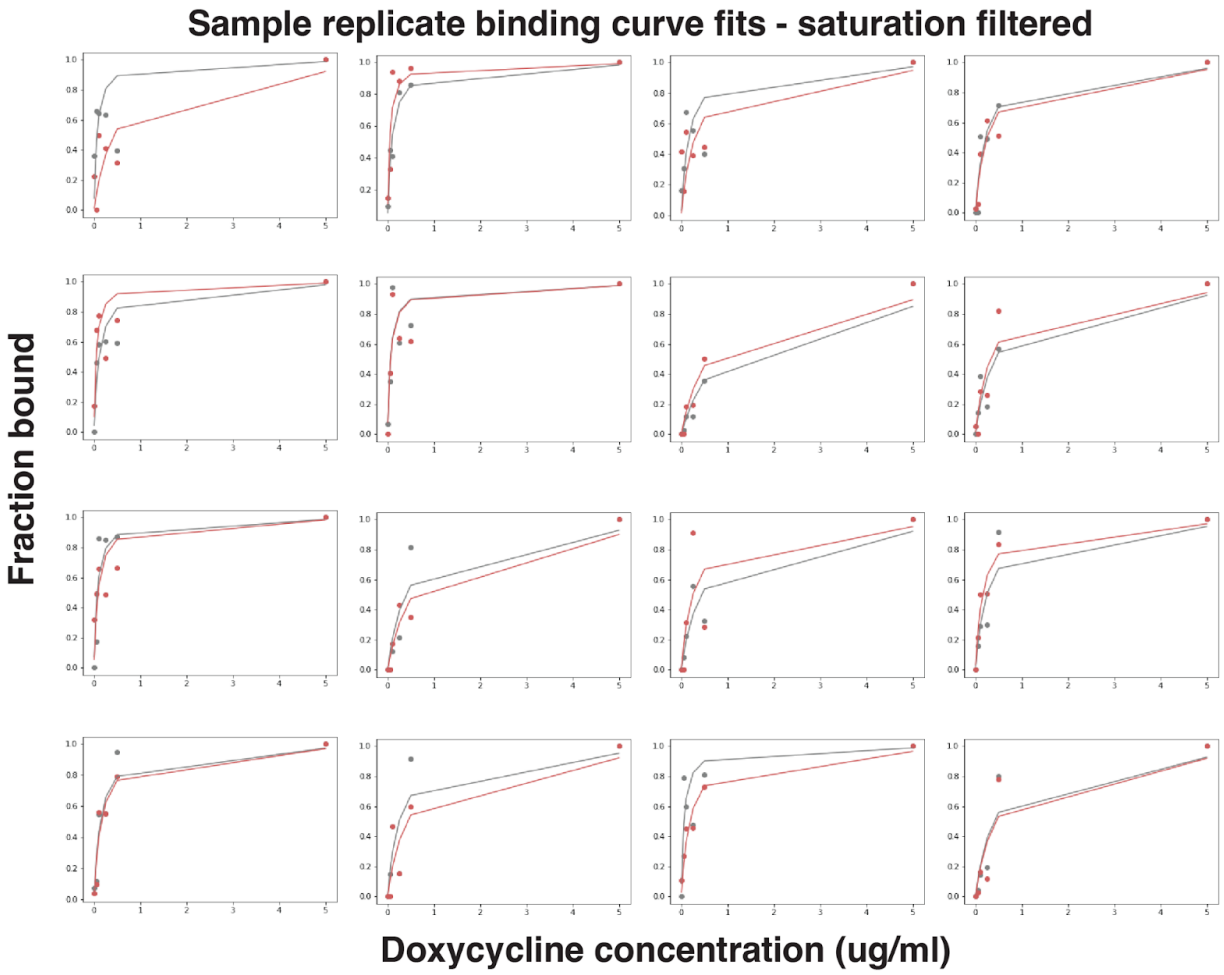

Figure 3.S3. Sample of replicate fit binding curves. RPKM binding signal and fitted lines for each CUT&Tag replicate at 16 representative genomic loci

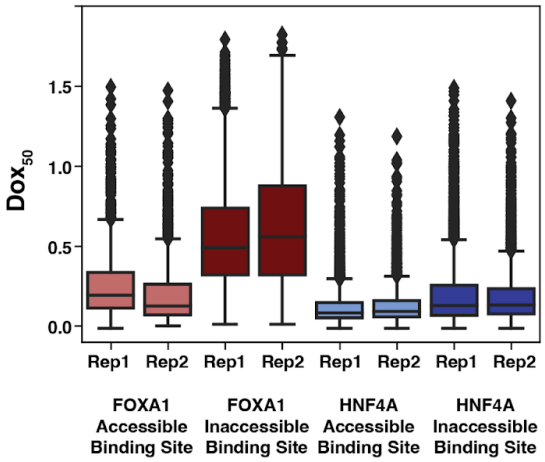

Figure 3.S4. Replicate dox$_{50}$ distributions. Dox$_{50}$ distributions extracted from fitted lines from each CUT&Tag replicate across each TF and each type of chromatin.

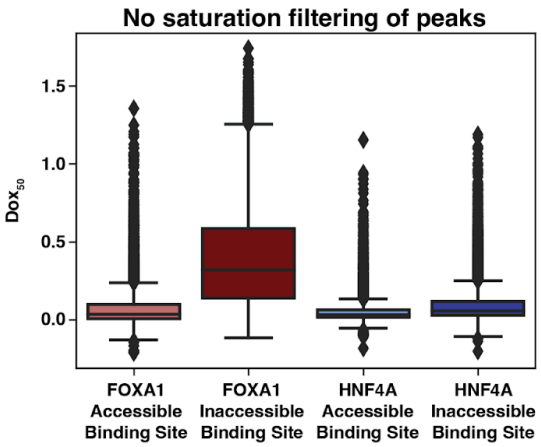

Figure 3.S5. Dox_{50} distributions without filtering out early saturation peaks. Dox50 distributions from all of the FOXA1 accessible binding sites (*n* = 10,118), FOXA1 inaccessible binding sites (*n* = 17,644), HNF4A accessible binding sites (*n* = 16,137), and HNF4A inaccessible binding sites (*n* = 16,507), without filtering out those peaks where binding signal peaked prior to the 5ug/ml dox sample.

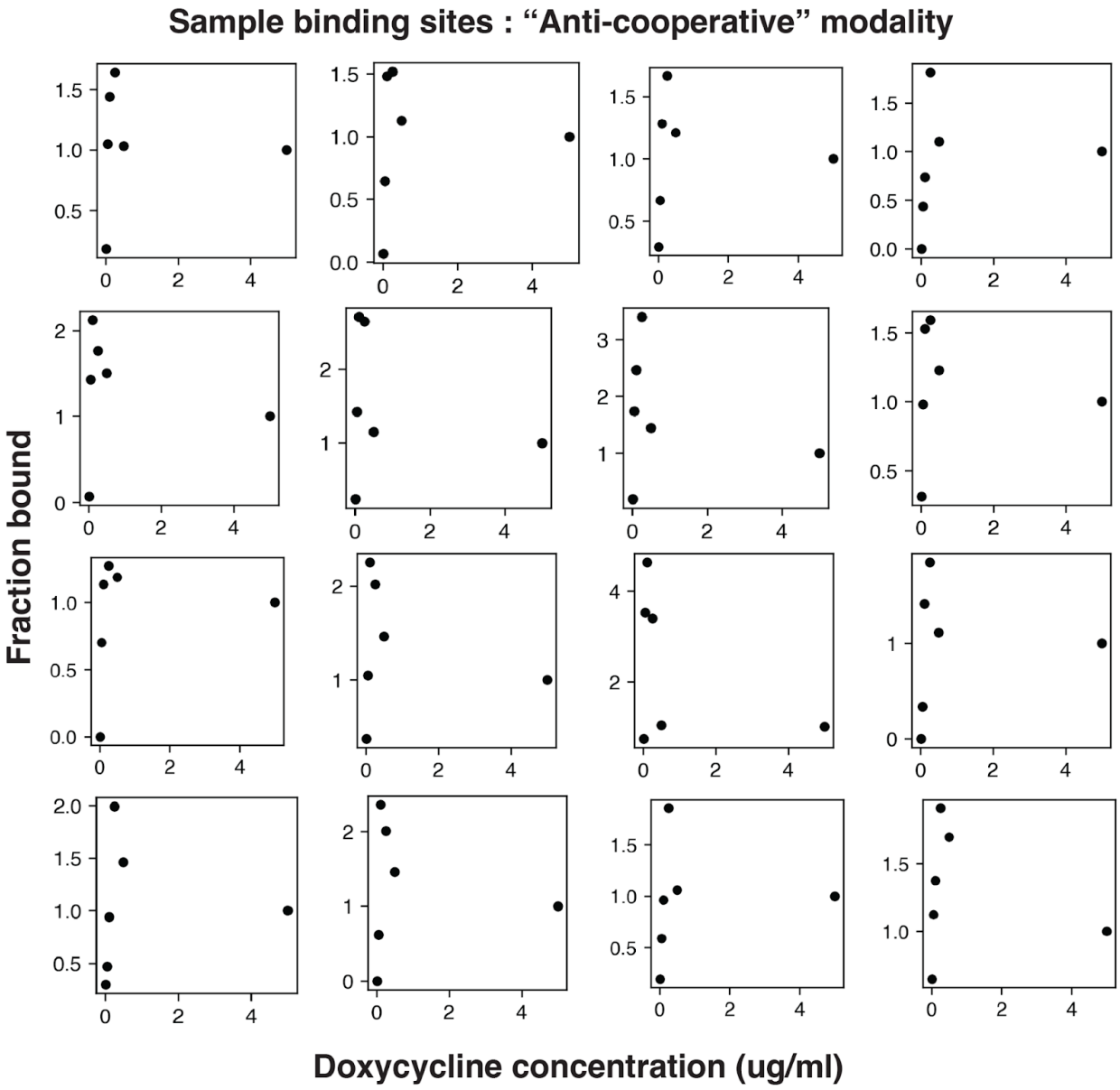

Figure 3.S6. Common “anti-cooperative" binding pattern. 16 examples from different genomic sites showing a pattern of increasing and then decreasing binding signal as dox induction increases. Signal is first read normalized (RPKM) and then normalized to the signal at the highest concentration. These sites were sampled from FOXA1 accessible binding sites.

Chapter 4 – Discussion

I designed the experiments and analyses throughout my thesis work to try to better understand how transcription factors (TFs) activate silent genes. I find this question interesting because of how it relates to complexity and diversity of human life. Skin cells look and behave differently from liver cells despite the fact that the two cell types contain nearly identical sets of genetic instructions. Seldom do skin cells spontaneously convert into liver cells. I discussed in Chapter 1 that cells prevent such spontaneous conversions by storing away genes from other lineages in a stable, inaccessible structure called heterochromatin. The heterochromatin sterically hinders TFs from accessing their regulatory targets.

And yet we discovered that we could reverse heterochromatic silencing and turn back on silent genes from other lineages by treating cells with a single TF from the desired ectopic lineage (Tapscott et al. 1988). Somehow the TF can find its targets within compacted DNA, decompact it, and then turn on nearby genes. We later developed strategies to leverage this phenomenon to convert cells back-in-time into their stem cell progenitors (Takahashi and Yamanaka 2006) and to convert cells directly across lineages into new cell types (Samantha A. Morris 2016). Despite these progresses though, each of these conversions is inefficient. Cells from the starting population often end up in "dead-end" states (Biddy et al. 2018), many genes necessary to the desired final cell type are never activated (Manandhar et al. 2017), and the final cells are often stuck in developmentally immature states (Biddy et al. 2018). While we know that TFs are capable of inducing these conversions, we do not fully understand how they are inducing these

conversions. I speculate that our incomplete understanding of silent gene activation is limiting our ability to realize cellular reprogramming's full potential.

The pioneer factor hypothesis (PFH) is currently the leading model for how TFs activate silent genes. It makes two claims. First, there are qualitatively different subsets of TFs: pioneer factors (PFs) can bind and open inaccessible chromatin and non-pioneer factors (nonPFs) cannot (Lisa Ann Cirillo et al. 2002; Iwafuchi-Doi and Zaret 2014). And second, activation of silent genes requires a two-step process in which PFs bind and open and then nonPFs are recruited to activate transcription (Lisa Ann Cirillo et al. 2002; Iwafuchi-Doi and Zaret 2014). Despite widespread adoption of the terminology, there is not a lot of direct evidence to support these two claims. This is especially true for the perhaps over-simplistic claim that there is a qualitative binary distinction between PFs and nonPFs.

One stark example in contradiction with this claim is the pioneer activity of the bacterial TF LEXA (Miller and Widom 2003). Bacteria do not contain nucleosomes and so LEXA never faced evolutionary pressure to develop the ability to bind heterochromatin. Maybe pioneer activity then is just an extension of how TFs bind to DNA. In another stark example, genome-wide binding analyses of canonical PFs PAX7 and FOXA1 showed many cases where the TFs did not bind to heterochromatic instances of their motifs (Donaghey et al. 2018; Mayran et al. 2018). Shouldn't the PFH have been scrapped at this point? Maybe we are waiting for more direct tests that provide us with more direct evidence. Many works that I have cited expressed multiple TFs at once, measured phenotypes in a genome-wide and correlation-based approach, or were limited to in vitro settings. Expressing multiple TFs complicates drawing conclusions about

single TFs, studying genome-wide phenotypes adds noise to the signal, and in vitro assays miss the complexity of the nucleus. Therefore in my work I aimed to design simple and direct tests of the PFH so that we could directly show whether the PFH's two claims are supported by the behavior of one canonical PF/nonPF pair. Once I showed that the claims were not supported, I used the same straightforward experimental design to propose an alternative model for silent gene activation.

4.1 – FOXA1 and HNF4A do not exhibit qualitatively different behavior

We chose FOXA1 and HNF4A as a canonical pair of PF and nonPF. FOXA1 is the most well studied PF (Lisa Ann Cirillo et al. 2002; Clark et al. 1993; Ramakrishnan et al. 1993; Donaghey et al. 2018) and FOXA1 and HNF4A have been suggested to behave as PF and nonPF in a reprogramming cocktail that converts fibroblasts into induced endoderm progenitors (Horisawa et al. 2020; Sekiya and Suzuki 2011; Biddy et al. 2018). Thus the PFH predicts that if we ectopically express FOXA1 and HNF4A, only FOXA1 will bind at inaccessible sites and neither TF will activate much tissue-specific gene expression. When we tested these predictions by individually expressing FOXA1 or HNF4A in K562 blood cells, we found that in fact both TFs bound and opened inaccessible sites and both TFs independently activated tissue-specific gene expression. In fact HNF4A activated more tissue-specific gene expression than FOXA1! We were surprised that these findings so starkly contradicted the PFH and so we searched for a distinction between the TFs' binding ability. While we found evidence suggesting that FOXA1 may be a stronger binder to naked DNA than HNF4A (Garcia et al. 2019; Jiang, Lee, and Sladek

1997; Rufibach et al. 2006), we showed in our data that HNF4A bound to more copies of its motif than FOXA1. This suggests that while HNF4A may bind more weakly to its site on naked DNA, it can nevertheless still achieve pioneer activity given the right sequence context. From these data we concluded that there is nothing qualitatively different about the pioneer activities of FOXA1 and HNF4A. Further, we suggested that pioneer activity is likely a general, quantitative quality of all TFs that depends on TF concentration, binding strength, and motif content at each target site. My work is only the latest to begin blurring the line between PFs and nonPFs: FOXA1 has been shown to rely on other TFs for pioneer activity (Swinstead et al. 2016) and HNF4A very recently has been shown to pioneer for the glucocorticoid receptor (Hunter et al. 2022).

As accumulating data further blur the line between PFs and nonPFs, there has been an accompanying unsuccessful attempt to identify some TF characteristic that could clearly distinguish the two types of TFs. If pioneer activity were a special trait limited to a subset of TFs, then there should be something structurally that we can find to be unique to these TFs. For example, proteins that can participate in phase separation are uniquely intrinsically disordered (Posey, Holehouse, and Pappu 2018). Semi-successful efforts have identified some helical motifs in the DNA-binding domains of some PFs, though the helical motifs seem neither necessary or sufficient for pioneer activity (Garcia et al. 2019). Instead, each TF uses a different strategy to achieve pioneer activity. Some bind exclusively at the entry/exit point of the histone core (Yu and Buck 2019) and others use partial motifs to access nucleosome-bound DNA (Soufi et al. 2015). These two behaviors alone suggest quite different mechanisms. In the former, the TF cannot bind to nucleosomal DNA and must capture its motif when it “breathes” away from the

nucleosomal edge (Polach and Widom 1995). In the latter, the TF has some structural binding domain that can still recognize its motif when the other side of the DNA is occluded by the nucleosomal core. And these are just two of the many other mechanisms used to bind to nucleosomes (F. Zhu et al. 2018), suggesting that many TFs independently evolved varying abilities to access their motifs within heterochromatin.

4.2 – HNF4A exhibits stronger quantitative pioneer activity than FOXA1

Having shown that FOXA1 and HNF4A do not behave qualitatively differently, we next aimed to establish a metric that could capture each TF's quantitative pioneer activity. The metric should capture a TF's binding strength at inaccessible DNA. We speculated that in vivo relative binding strengths would be appropriate. In vivo measurements provide physiological relevance and relative binding strength measurements control for any differences between our TF-expression lines. To capture binding strength, we used our doxycycline (dox) induction system to make a K_d-like measurement that we called dox_{50}. A TF's dox_{50} is the concentration of dox required for the TF to half-maximally bind a certain genomic site. And to capture a TF's relative binding strength between inaccessible and accessible sites, we took the average difference between a TF's inaccessible and accessible dox_{50}s. We called this a TF's Δdox_{50}. A low Δdox_{50} indicates strong pioneer activity. When we induced each TF across a 1,000-fold range of dox, measured binding, fit curves, and extracted dox_{50}s, we found that HNF4A has a lower Δdox_{50} and thus stronger pioneer activity than FOXA1. We also found that FOXA1 had stronger pioneer activity at sites where it could target more copies of its motif, suggesting that strong motif content could

somewhat make up for lower pioneer activity. From these data we propose that all TFs likely have some degree of pioneer activity and that a TF's Δdox_{50} (or some related metric) could eventually become just another way of quantifying TF behavior.

Like our data refuting qualitative pioneer activity, this data is not altogether unexpected. If pioneer activity is quantitative, then there should be dials that can be turned in order to increase or decrease it. These dials are likely the parameters that drive high TF occupancy: TF concentration, binding strength, and motif content. There is good evidence that shows that these dials are important. In one study, a group was able to endow nonPFs with pioneer activity by increasing the concentration of the nonPFs (Yan, Chen, and Bai 2018). Similarly, when we reduced the concentration of FOXA1 by reducing the dox induction level, we found that while its overall binding signal dropped genome-wide, this decrease was much more pronounced at inaccessible binding sites (Hansen, Loell, and Cohen 2022). Motif content also appears important. In one study, the quality of the motif impacted whether or not the same TF behaved as a PF or nonPF (Meers, Janssens, and Henikoff 2019). And in my work, HNF4A used more motifs than FOXA1 in order to achieve pioneer activity (Hansen, Loell, and Cohen 2022). Perhaps our traditional PFs were all classified this way within regimes where they were highly expressed or where they targeted sites with strong motif content. If we expand our view to assume that all TFs have some pioneer activity, then we can more holistically study how TFs interact with nucleosomal DNA and more widely select TFs for use in future ectopic gene activation assays.

4.3 – Why did the PFH have it wrong?

We were initially surprised by our first qPCR experiments that showed that ectopically expressed HNF4A strongly activated a couple of liver genes in K562 blood cells. It is striking alone that any TF could turn on heterochromatically-silenced liver genes in a blood cancer line. It is further striking that a nonPF could do it. Perhaps though these genes had enhancers in accessible regions of the K562 genome, making them easier for HNF4A to access. But then we showed that like FOXA1, HNF4A activated many liver- and intestine-specific genes in K562 blood cells and did so by binding to many genome-wide inaccessible sites. Collectively these data made us question why a binary distinction was ever drawn between FOXA1 and HNF4A. The data also made us question what it really means when we label DNA as "accessible" or "inaccessible."

The first reason a distinction may have been drawn between FOXA1 and HNF4A is that there is some evidence that FOXA1 binds more strongly to naked DNA than HNF4A (Garcia et al. 2019; Jiang, Lee, and Sladek 1997; Rufibach et al. 2006). It is unclear whether this sort of measured binding strength can be appropriately applied to in vivo chromatin contexts, but it would suggest that if each TF were presented with a motif, FOXA1 could bind it more easily. Further, FOXA1 is known to have a DNA-binding domain with a similar three-dimensional structure to the linker histone protein H1, more evidence that FOXA1 may be more proficient at in vivo DNA binding (Clark et al. 1993; Ramakrishnan et al. 1993). On the other hand, HNF4A may be able to compensate for weaker binding by forming dimers with other HNF4A molecules. HNF4A is known to bind either as a homodimer or as a heterodimer with isomers of itself. (Jiang et al. 1995; Ko, Zhuo, and Ren 2019) and we find that HNF4A has multiple copies of its motifs at sites

where it binds inaccessible DNA (Hansen, Loell, and Cohen 2022). Dimerized proteins are larger protein complexes that recognize twice as much DNA sequence and therefore may use their size and specificity to outcompete histones. CAS9 is similarly large, recognizes a long RNA-guided recognition sequence, and has been shown to exhibit pioneer activity (Barkal et al. 2016). HNF4A's reliance on dimerization for pioneer activity could be tested further by manipulating its dimerization domain and then re-measuring its ability to bind at inaccessible motifs.

Developmental importance may also have contributed to FOXA1 and HNF4A's classification, though in my opinion each TF's role and timing through development are not entirely clear. While it does appear that FOXA1 is critical for the development of the liver bud (Lee et al. 2005) and that HNF4A is critical for gastrulation (Chen et al. 1994), knocking out HNF4A in fetal mice disrupts liver architecture (Parviz et al. 2003). This suggests that HNF4A may also play an important role in liver development. It's also possible that HNF4A may actually precede FOXA1 in embryonic liver development by half a day (Lau et al. 2018). If this is true, then perhaps we should have labeled HNF4A as the PF and expressed it prior to FOXA1 during reprogramming experiments.

And finally I caution over-interpreting the label of "inaccessible DNA." The most common technique used these days to measure DNA accessibility is ATAC-sequencing (Buenrostro et al. 2015), in which DNA is exposed to a sequencing primer-laden transposase enzyme so that only DNA that was "attacked" by the enzyme is sequenced. The regions where the transposase can access are thus "accessible," and vice versa. But just because a region is inaccessible to transposase doesn't mean that it is inaccessible to a TF. A transposase is an enzyme with no

sequence specificity whereas a TF is guided to its targets by strong protein-DNA interactions; the latter event seems more likely to occur than the former. And even if we assume that transposases and TFs have similar binding abilities, we know that nucleosomal DNA is not static but rather can "breathe" away from the histones and create transiently free DNA (Polach and Widom 1995). Perhaps we should consider the follow-up question "to what?" when DNA is classified as inaccessible.

4.4 – Incomplete silent gene activation limits cellular reprogramming

Our incomplete understanding of how HNF4A behaves when individually expressed in an ectopic setting suggests that we may not have been using it correctly in reprogramming experiments. This may be one factor that explains why the FOXA1-HNF4A reprogramming cocktail produces either "dead-end cells" or developmentally immature endoderm progenitors (Biddy et al. 2018). I suspect that other cocktails too are limited by an incomplete understanding of how the individual TFs behave. There are several ways that we could use what I've learned about HNF4A to improve the conversion to a more purely liver-life final state. The first is the order of expression and the second is the level of expression.

As mentioned earlier, there may be a specific sequence of TF activities in the developing endoderm that requires first the activity of one TF and then the activity of another. Currently the FOXA1-HNF4A reprogramming cocktail simultaneously expresses both TFs. Perhaps HNF4A needs to set up an endoderm-specific nucleus after which FOXA1 can push the cell towards the

liver lineage. In this case we should use constitutive HNF4A expression and inducible FOXA1. Or perhaps FOXA1 needs to lock the converting cells into the liver lineage after which HNF4A can help turn on some additional liver-specific genes. In this case we would use constitutive FOXA1 and inducible HNF4A.

The current cocktail also might be combining two PFs for different lineages to produce a monster hybrid cell. FOXA1 is pioneering liver genes and HNF4A is pioneering intestine genes and when neither wins out, the resultant cells either end up in a developmentally immature endoderm cell or just lack the appropriate feedback signals to proceed. Currently the two TFs are expressed at essentially identical levels. It would be interesting to use an inducible HNF4A construct to test a dose-response curve to see if there is an optimal HNF4A concentration at which the intestine lineage disappears but where HNF4A still turns on some liver genes. I intended to test both timing and dosage of the two TFs and cloned a construct with constitutive FOXA1 and inducible HNF4A, but unfortunately ran out of time. I passed the construct to the Morris lab, who regularly performs these conversions and who perhaps may incorporate this new construct and/or other modifications into future reprograming cocktails (Biddy et al. 2018).

I would have also liked to test additional PF/nonPF pairs. I think it would be valuable to dissect other reprogramming cocktails as I have dissected the FOXA1-HNF4A cocktail to better understand what each TF can do individually. Ideally, we could test many TFs from many lineages in order to build a catalog of reprogramming TFs (see below) but a good start would include those TFs from prominent cocktails. The next cocktail that I had in mind combines PF ASCL1 and nonPF MYT1L to convert fibroblasts into neurons (Vierbuchen et al. 2010;

Wapinski et al. 2013; Treutlein et al. 2016). As I mentioned in Chapter 1, this sort of reprogramming could someday be used to replace neurons lost to stroke or other brain damage. But if we hope to use the converted neurons in humans, then we need to be certain that we are generating the intended cell type and thus must turn on the appropriate genes. In my final section I recommend a gene regulation-based strategy for designing future reprogramming cocktails.

4.5 – Rational design of new reprogramming cocktails

To date our reprogramming cocktails have mostly been built by testing many TFs for their ability to create some sort of meaningful phenotype like colony formation (Biddy et al. 2018) or drug oxidation (Sekiya and Suzuki 2011). We expect TFs do this by activating silent genes but we rarely measure the silent gene activation itself. This approach has led to the misclassification of some TFs as incapable of governing conversion processes and in doing so may have slowed our progress towards more successful reprogramming. There is quite exciting potential if we are able to efficiently produce new cardiomyocytes to replace those lost in a heart attack but we need to ensure that the process of creating them is fully understood and carefully controlled. I hope that my work could serve as one example for how a reprogramming cocktail could be tested or conceived. Based on my experiments I recommend two steps.

First, we should develop TF induction systems across multiple (or many) cell types and then measure which genes are activated. I speculate that the overlap in activated genes between repeated experiments will inform on the “true targets” of a given TF. I predict that the stronger a TF’s activity, the fewer cell types will need to be tested to find this set. Once we have lists of

gene targets for each TF, then whenever down the road we encounter a need to activate an individual or set of genes, we would know the appropriate TF(s). I imagine that we could similarly collect a set of true binding sites for different TFs. In experiments where TFs are ectopically activated, conclusions are often drawn from the genome-wide patterns of tens of thousands of binding events (Donaghey et al. 2018). If we consider that the main role TFs play is activating genes by binding to enhancers, and there are approximately 200 tissue-specific genes per lineage (Uhlén et al. 2015), and between 5-10 enhancers per gene (Fishilevich et al. 2017; Andersson et al. 2014), then we would expect on the order of 1000-2000 functional binding sites, not tens of thousands. If we were able to limit our analysis of TF binding to this smaller set of sites, then perhaps we could use a higher signal to noise ratio to identify important sequence features for binding.

Once we have a set of targets for each TF, then I would recommend that we measure a Δdox_{50} (or similar metric) for each TF to understand the TF's pioneer activity. This information as well as knowing the concentration at which the TF is expressed and the motifs surrounding each of the targets would provide us the three aforementioned dials that we can turn to control pioneer activity. For instance, if we find that a TF has a high Δdox_{50} (or weak pioneer activity), then we could induce the TF at higher concentrations to make sure that its target genes are activated. And if this is not possible, then we could at least predict that perhaps only the targets that have strong nearby motif content would be activated, in line with our finding that strong motif content could boost pioneer activity. Finally, having this metric would allow us to conduct a more thorough analysis of what structural motifs might relate to pioneer activity by correlating TFs' Δdox_{50} with other quantitative metrics, like size.

4.6 – Conclusion

I hope that my work added a little bit of clarity to how we think about the mechanism by which TFs activate silent genes. Many of the experiments that I cited had complicated designs and used complicated figures to argue that there is a binary classification between PFs and nonPFs. I too initially designed and tried to implement a complicated experimental system to study which sequences could direct pioneer activity. Sometimes these complex experiments can be effective at allowing us to test many elements in parallel or in ultra-controlled environments and so we should not abandon them. But we should also remember that if we can simplify the question that we are asking into a hypothesis and straightforward predictions, then a simple experimental system may be sufficient.

After my initial experiments did not pan out (and I was desperate for the subsequent ones to work) I went back to the PFH and realized that it made two predictions that I was entirely capable of testing. First, ectopic HNF4A should not bind inaccessible sites. And second, neither FOXA1 nor HNF4A should activate much liver- or intestine-specific gene expression. It was easy to learn how to perform lentiviral transductions, RNA-sequencing, ATAC-sequencing, and CUT&Tag; these are all broadly utilized techniques. Upon implementing them, I showed that both predictions of the PFH were wrong. At least for FOXA1 and HNF4A, these data should kill the PFH! I was then able to extend the same experimental system to induce FOXA1 and HNF4A across a wide dynamic range and develop a quantitative metric for pioneer activity.

There are a lot of experiments that I wish that I had remaining time to perform. I would like to test various time courses and dose responses of FOXA1 and HNF4A in an attempt to better reprogram fibroblasts, I would like to dissect additional reprogramming cocktails to study the individual TFs' pioneer activities, I would like to express FOXA1 and HNF4A in another cell line to move closer to each TF's list of "true gene targets," and I would like to further test how HNF4A might be exhibiting pioneer activity, perhaps by manipulating its dimerization domain. Maybe lab mates of mine or readers of my work will seize the torch! In any case, I'm excited to see how my data and whatever comes next may allow us to build more efficient reprogramming cocktails from a stronger foundation of silent gene activation. And I'm eager to learn how the collective work of my lab, myself, and the field move us closer to understanding how the enormously complex sequence of the human genome can encode for such a cool thing as life.

References

Andersson, Robin, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, et al. 2014. “An Atlas of Active Enhancers across Human Cell Types and Tissues.” *Nature* 507 (7493): 455–61.

Bailey, Timothy L. 2021. “STREME: Accurate and Versatile Sequence Motif Discovery.” *Bioinformatics* , March. https://doi.org/10.1093/bioinformatics/btab203.

Barkal, Amira A., Sharanya Srinivasan, Tatsunori Hashimoto, David K. Gifford, and Richard I. Sherwood. 2016. “Cas9 Functionally Opens Chromatin.” *PloS One* 11 (3): e0152683.

Barozzi, Iros, Marta Simonatto, Silvia Bonifacio, Lin Yang, Remo Rohs, Serena Ghisletti, and Gioacchino Natoli. 2014. “Coregulation of Transcription Factor Binding and Nucleosome Occupancy through DNA Features of Mammalian Enhancers.” *Molecular Cell* 54 (5): 844–57.

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.

Biddy, Brent A., Wenjun Kong, Kenji Kamimoto, Chuner Guo, Sarah E. Waye, Tao Sun, and Samantha A. Morris. 2018. “Single-Cell Mapping of Lineage and Identity in Direct Reprogramming.” *Nature* 564 (7735): 219–24.

Boller, Sören, Senthilkumar Ramamoorthy, Duygu Akbas, Robert Nechanitzky, Lukas Burger, Rabih Murr, Dirk Schübeler, and Rudolf Grosschedl. 2016. “Pioneering Activity of the C-Terminal Domain of EBF1 Shapes the Chromatin Landscape for B Cell Programming.” *Immunity* 44 (3): 527–41.

Boyes, Joan, and Gary Felsenfeld. 1996. “Tissue-Specific Factors Additively Increase the Probability of the All-or-None Formation of a Hypersensitive Site.” *The EMBO Journal* 15 (10): 2496–2507.

Buenrostro, Jason D., Beijing Wu, Howard Y. Chang, and William J. Greenleaf. 2015. “ATAC-Seq: A Method for Assaying Chromatin Accessibility Genome-Wide.” *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* 109 (January): 21.29.1–9.

Casey, Bradford H., Rahul K. Kollipara, Karine Pozo, and Jane E. Johnson. 2018. “Intrinsic DNA Binding Properties Demonstrated for Lineage-Specifying Basic Helix-Loop-Helix Transcription Factors.” *Genome Research* 28 (4): 484–96.

Chang, Yujung, Euiyeon Lee, Junyeop Kim, Yoo-Wook Kwon, Youngeun Kwon, and Jongpil Kim. 2019. “Efficient in Vivo Direct Conversion of Fibroblasts into Cardiomyocytes Using a Nanoparticle-Based Gene Carrier.” *Biomaterials* 192 (February): 500–509.

Chen, W. S., K. Manova, D. C. Weinstein, S. A. Duncan, A. S. Plump, V. R. Prezioso, R. F. Bachvarova, and J. E. Darnell Jr. 1994. “Disruption of the HNF-4 Gene, Expressed in Visceral Endoderm, Leads to Cell Death in Embryonic Ectoderm and Impaired Gastrulation of Mouse Embryos.” *Genes & Development* 8 (20): 2466–77.

Choi, J., M. L. Costa, C. S. Mermelstein, C. Chagas, S. Holtzer, and H. Holtzer. 1990. “MyoD Converts Primary Dermal Fibroblasts, Chondroblasts, Smooth Muscle, and Retinal Pigmented Epithelial Cells into Striated Mononucleated Myoblasts and Multinucleated Myotubes.” *Proceedings of the National Academy of Sciences* 87 (20): 7988–92.

Cirillo, L. A., C. E. McPherson, P. Bossard, K. Stevens, S. Cherian, E. Y. Shim, K. L. Clark, S. K. Burley, and K. S. Zaret. 1998. “Binding of the Winged-Helix Transcription Factor HNF3 to a Linker Histone Site on the Nucleosome.” *The EMBO Journal* 17 (1): 244–54.

Cirillo, Lisa Ann, Frank Robert Lin, Isabel Cuesta, Dara Friedman, Michal Jarnik, and Kenneth S. Zaret. 2002. “Opening of Compacted Chromatin by Early Developmental Transcription Factors HNF3 (FoxA) and GATA-4.” *Molecular Cell* 9 (2): 279–89.

Clark, K. L., E. D. Halay, E. Lai, and S. K. Burley. 1993. “Co-Crystal Structure of the HNF-3/fork Head DNA-Recognition Motif Resembles Histone H5.” *Nature* 364 (6436): 412–20.

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.

Davis, R. L., H. Weintraub, and A. B. Lassar. 1987. “Expression of a Single Transfected cDNA Converts Fibroblasts to Myoblasts.” *Cell* 51 (6): 987–1000.

Donaghey, Julie, Sudhir Thakurela, Jocelyn Charlton, Jennifer S. Chen, Zachary D. Smith, Hongcang Gu, Ramona Pop, et al. 2018. “Genetic Determinants and Epigenetic Effects of Pioneer-Factor Occupancy.” *Nature Genetics* 50 (2): 250–58.

Elgin, S. C. 1996. “Heterochromatin and Gene Regulation in Drosophila.” *Current Opinion in Genetics & Development* 6 (2): 193–202.

Elkon, Ran, and Reuven Agami. 2017. “Characterization of Noncoding Regulatory DNA in the Human Genome.” *Nature Biotechnology* 35 (8): 732–46.

Ernst, Jason, and Manolis Kellis. 2012. “ChromHMM: Automating Chromatin-State Discovery and Characterization.” *Nature Methods* 9 (February): 215.

Fishilevich, Simon, Ron Nudel, Noa Rappaport, Rotem Hadar, Inbar Plaschkes, Tsippi Iny Stein, Naomi Rosen, et al. 2017. “GeneHancer: Genome-Wide Integration of Enhancers and Target Genes in GeneCards.” *Database: The Journal of Biological Databases and Curation* 2017 (January). https://doi.org/10.1093/database/bax028.

Fornes, Oriol, Jaime A. Castro-Mondragon, Aziz Khan, Robin van der Lee, Xi Zhang, Phillip A. Richmond, Bhavi P. Modi, et al. 2020. “JASPAR 2020: Update of the Open-Access Database of Transcription Factor Binding Profiles.” *Nucleic Acids Research* 48 (D1): D87–92.

Furuyama, Kenichiro, Simona Chera, Léon van Gurp, Daniel Oropeza, Luiza Ghila, Nicolas Damond, Heidrun Vethe, et al. 2019. “Diabetes Relief in Mice by Glucose-Sensing Insulin-Secreting Human α-Cells.” *Nature*, February. https://doi.org/10.1038/s41586-019-0942-8.

Garcia, Meilin Fernandez, Cedric D. Moore, Katharine N. Schulz, Oscar Alberto, Greg Donague, Melissa M. Harrison, Heng Zhu, and Kenneth S. Zaret. 2019. “Structural Features of Transcription Factors Associating with Nucleosome Binding.” *Molecular Cell*. https://doi.org/10.1016/j.molcel.2019.06.009.

Grant, Charles E., and Timothy L. Bailey. 2021. “XSTREME: Comprehensive Motif Analysis of Biological Sequence Datasets.” *bioRxiv*. https://doi.org/10.1101/2021.09.02.458722.

Grant, Charles E., Timothy L. Bailey, and William Stafford Noble. 2011. “FIMO: Scanning for Occurrences of a given Motif.” *Bioinformatics* 27 (7): 1017–18.

Gurdon, J. B. 1962. “Adult Frogs Derived from the Nuclei of Single Somatic Cells.” *Developmental Biology* 4 (April): 256–73.

Hammelman, Jennifer, Konstantin Krismer, Budhaditya Banerjee, David K. Gifford, and Richard I. Sherwood. 2020. “Identification of Determinants of Differential Chromatin Accessibility through a Massively Parallel Genome-Integrated Reporter Assay.” *Genome Research* 30 (10): 1468–80.

Hansen, Jeffrey L., Kaiser J. Loell, and Barak A. Cohen. 2022. “The Pioneer Factor Hypothesis Is Not Necessary to Explain Ectopic Liver Gene Activation.” *eLife* 11 (January). https://doi.org/10.7554/eLife.73358.

Heintzman, Nathaniel D., Gary C. Hon, R. David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F. Harp, Zhen Ye, et al. 2009. “Histone Modifications at Human Enhancers Reflect Global Cell-Type-Specific Gene Expression.” *Nature* 459 (7243): 108–12.

Heintzman, Nathaniel D., Rhona K. Stuart, Gary Hon, Yutao Fu, Christina W. Ching, R. David Hawkins, Leah O. Barrera, et al. 2007. “Distinct and Predictive Chromatin Signatures of Transcriptional Promoters and Enhancers in the Human Genome.” *Nature Genetics* 39 (3): 311–18.

Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 2010. "Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities." *Molecular Cell* 38 (4): 576–89.

Heitz, Emil. 1928. *Das Heterochromatin Der Moose*. Bornträger.

Horisawa, Kenichi, Miyako Udono, Kazuko Ueno, Yasuyuki Ohkawa, Masao Nagasaki, Sayaka Sekiya, and Atsushi Suzuki. 2020. "The Dynamics of Transcriptional Activation by Hepatic Reprogramming Factors." *Molecular Cell* 79 (4): 660–76.e8.

Huertas, Jan, Caitlin M. MacCarthy, Hans R. Schöler, and Vlad Cojocaru. 2020. "Nucleosomal DNA Dynamics Mediate Oct4 Pioneer Factor Binding." *Biophysical Journal*, January. https://doi.org/10.1016/j.bpj.2019.12.038.

Huh, Christine J., Bo Zhang, Matheus B. Victor, Sonika Dahiya, Luis Fz Batista, Steve Horvath, and Andrew S. Yoo. 2016. "Maintenance of Age in Human Neurons Generated by microRNA-Based Neuronal Conversion of Fibroblasts." *eLife* 5 (September). https://doi.org/10.7554/eLife.18648.

Hunter, A. Louise, Toryn M. Poolman, Donghwan Kim, Frank J. Gonzalez, David A. Bechtold, Andrew S. I. Loudon, Mudassar Iqbal, and David W. Ray. 2022. "HNF4A Modulates Glucocorticoid Action in the Liver." *Cell Reports* 39 (3): 110697.

Ieda, Masaki, Ji-Dong Fu, Paul Delgado-Olguin, Vasanth Vedantham, Yohei Hayashi, Benoit G. Bruneau, and Deepak Srivastava. 2010. "Direct Reprogramming of Fibroblasts into Functional Cardiomyocytes by Defined Factors." *Cell* 142 (3): 375–86.

Iwafuchi-Doi, Makiko, and Kenneth S. Zaret. 2014. "Pioneer Transcription Factors in Cell Reprogramming." *Genes & Development* 28 (24): 2679–92.

Jayawardena, Tilanthi M., Elizabeth A. Finch, Lunan Zhang, Hengtao Zhang, Conrad P. Hodgkinson, Richard E. Pratt, Paul B. Rosenberg, Maria Mirotsou, and Victor J. Dzau. 2015. "MicroRNA Induced Cardiac Reprogramming in Vivo: Evidence for Mature Cardiac Myocytes and Improved Cardiac Function." *Circulation Research* 116 (3): 418–24.

Jiang, G., U. Lee, and F. M. Sladek. 1997. "Proposed Mechanism for the Stabilization of Nuclear Receptor DNA Binding via Protein Dimerization." *Molecular and Cellular Biology* 17 (11): 6546–54.

Jiang, G., L. Nepomuceno, K. Hopkins, and F. M. Sladek. 1995. "Exclusive Homodimerization of the Orphan Receptor Hepatocyte Nuclear Factor 4 Defines a New Subclass of Nuclear Receptors." *Molecular and Cellular Biology* 15 (9): 5131–43.

Jopling, Chris, Eduard Sleep, Marina Raya, Mercè Martí, Angel Raya, and Juan Carlos Izpisúa Belmonte. 2010. “Zebrafish Heart Regeneration Occurs by Cardiomyocyte Dedifferentiation and Proliferation.” *Nature* 464 (7288): 606–9.

Kaplan, Noam, Irene K. Moore, Yvonne Fondufe-Mittendorf, Andrea J. Gossett, Desiree Tillo, Yair Field, Emily M. LeProust, et al. 2009. “The DNA-Encoded Nucleosome Organization of a Eukaryotic Genome.” *Nature* 458 (7236): 362–66.

Karagianni, Panagiota, Panagiotis Moulos, Dominic Schmidt, Duncan T. Odom, and Iannis Talianidis. 2020. “Bookmarking by Non-Pioneer Transcription Factors during Liver Development Establishes Competence for Future Gene Activation.” *Cell Reports* 30 (5): 1319–28.e6.

Kaya-Okur, Hatice S., Derek H. Janssens, Jorja G. Henikoff, Kami Ahmad, and Steven Henikoff. 2020. “Efficient Low-Cost Chromatin Profiling with CUT&Tag.” *Nature Protocols* 15 (10): 3264–83.

Kaya-Okur, Hatice S., Steven J. Wu, Christine A. Codomo, Erica S. Pledger, Terri D. Bryson, Jorja G. Henikoff, Kami Ahmad, and Steven Henikoff. 2019. “CUT&Tag for Efficient Epigenomic Profiling of Small Samples and Single Cells.” *Nature Communications* 10 (1): 1930.

Ko, Hui Ling, Ziyi Zhuo, and Ee Chee Ren. 2019. “HNF4α Combinatorial Isoform Heterodimers Activate Distinct Gene Targets That Differ from Their Corresponding Homodimers.” *Cell Reports* 26 (10): 2549–57.e3.

Kornberg, R. D. 1974. “Chromatin Structure: A Repeating Unit of Histones and DNA.” *Science* 184 (4139): 868–71.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.

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.

Larson, Elizabeth D., Hideyuki Komori, Tyler J. Gibson, Cyrina M. Ostgaard, Danielle C. Hamm, Jack M. Schnell, Cheng-Yu Lee, and Melissa M. Harrison. 2021. “Cell-Type-Specific Chromatin Occupancy by the Pioneer Factor Zelda Drives Key Developmental Transitions in Drosophila.” *Nature Communications* 12 (1): 7153.

Lau, Hwee Hui, Natasha Hui Jin Ng, Larry Sai Weng Loo, Joanita Binte Jasmen, and Adrian Kee Keong Teo. 2018. “The Molecular Functions of Hepatocyte Nuclear Factors – In and beyond the Liver.” *Journal of Hepatology* 68 (5): 1033–48.

Lee, Catherine S., Joshua R. Friedman, James T. Fulmer, and Klaus H. Kaestner. 2005. “The Initiation of Liver Development Is Dependent on Foxa Transcription Factors.” *Nature* 435 (7044): 944–47.

Lemma, Roza B., Marit Ledsaak, Bettina M. Fuglerud, Geir Kjetil Sandve, Ragnhild Eskeland, and Odd S. Gabrielsen. 2021. “Chromatin Occupancy and Target Genes of the Haematopoietic Master Transcription Factor MYB.” *Scientific Reports* 11 (1): 9008.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. “featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics* 30 (7): 923–30.

Li, Qunhua, James B. Brown, Haiyan Huang, and Peter J. Bickel. 2011. “Measuring Reproducibility of High-Throughput Experiments.” *The Annals of Applied Statistics* 5 (3): 1752–79.

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.

Lupien, Mathieu, Jérôme Eeckhoute, Clifford A. Meyer, Qianben Wang, Yong Zhang, Wei Li, Jason S. Carroll, X. Shirley Liu, and Myles Brown. 2008. “FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-Specific Transcription.” *Cell* 132 (6): 958–70.

Ma, Hong, Li Wang, Chaoying Yin, Jiandong Liu, and Li Qian. 2015. “In Vivo Cardiac Reprogramming Using an Optimal Single Polycistronic Construct.” *Cardiovascular Research* 108 (2): 217–19.

Manandhar, Dinesh, Lingyun Song, Ami Kabadi, Jennifer B. Kwon, Lee E. Edsall, Melanie Ehrlich, Koji Tsumagari, Charles A. Gersbach, Gregory E. Crawford, and Raluca Gordân. 2017. “Incomplete MyoD-Induced Transdifferentiation Is Associated with Chromatin Remodeling Deficiencies.” *Nucleic Acids Research* 45 (20): 11684–99.

Man, T. K., and G. D. Stormo. 2001. “Non-Independence of Mnt Repressor-Operator Interaction Determined by a New Quantitative Multiple Fluorescence Relative Affinity (QuMFRA) Assay.” *Nucleic Acids Research* 29 (12): 2471–78.

Matsuda, Taito, Takashi Irie, Shutaro Katsurabayashi, Yoshinori Hayashi, Tatsuya Nagai, Nobuhiko Hamazaki, Aliya Mari D. Adefuin, et al. 2018. “Pioneer Factor NeuroD1 Rearranges Transcriptional and Epigenetic Profiles to Execute Microglia-Neuron Conversion.” *Neuron*, December. https://doi.org/10.1016/j.neuron.2018.12.010.

Mayran, Alexandre, Konstantin Khetchoumian, Fadi Hariri, Tomi Pastinen, Yves Gauthier, Aurelio Balsalobre, and Jacques Drouin. 2018. “Pioneer Factor Pax7 Deploys a Stable Enhancer Repertoire for Specification of Cell Fate.” *Nature Genetics* 50 (2): 259–69.

McDaniel, Stephen L., Tyler J. Gibson, Katharine N. Schulz, Meilin Fernandez Garcia, Markus Nevil, Siddhant U. Jain, Peter W. Lewis, Kenneth S. Zaret, and Melissa M. Harrison.

2019. “Continued Activity of the Pioneer Factor Zelda Is Required to Drive Zygotic Genome Activation.” *Molecular Cell*, February. https://doi.org/10.1016/j.molcel.2019.01.014.

McPherson, C. E., E. Y. Shim, D. S. Friedman, and K. S. Zaret. 1993. “An Active Tissue-Specific Enhancer and Bound Transcription Factors Existing in a Precisely Positioned Nucleosomal Array.” *Cell* 75 (2): 387–98.

Meerbrey, Kristen L., Guang Hu, Jessica D. Kessler, Kevin Roarty, Mamie Z. Li, Justin E. Fang, Jason I. Herschkowitz, et al. 2011. “The pINDUCER Lentiviral Toolkit for Inducible RNA Interference in Vitro and in Vivo.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (9): 3665–70.

Meers, Michael P., Derek H. Janssens, and Steven Henikoff. 2019. “Pioneer Factor-Nucleosome Binding Events during Differentiation Are Motif Encoded.” *Molecular Cell*, June. https://doi.org/10.1016/j.molcel.2019.05.025.

Miller, Joanna A., and Jonathan Widom. 2003. “Collaborative Competition Mechanism for Gene Activation in Vivo.” *Molecular and Cellular Biology* 23 (5): 1623–32.

Minderjahn, Julia, Andreas Schmidt, Andreas Fuchs, Rudolf Schill, Johanna Raithel, Magda Babina, Christian Schmidl, et al. 2020. “Mechanisms Governing the Pioneering and Redistribution Capabilities of the Non-Classical Pioneer PU.1.” *Nature Communications* 11 (1): 402.

Mirny, Leonid A. 2010. “Nucleosome-Mediated Cooperativity between Transcription Factors.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (52): 22534–39.

Morris, Samantha A. 2016. “Direct Lineage Reprogramming via Pioneer Factors; a Detour through Developmental Gene Regulatory Networks.” *Development* 143 (15): 2696–2705.

Morris, Samantha A., Patrick Cahan, Hu Li, Anna M. Zhao, Adrianna K. San Roman, Ramesh A. Shivdasani, James J. Collins, and George Q. Daley. 2014. “Dissecting Engineered Cell Types and Enhancing Cell Fate Conversion via CellNet.” *Cell* 158 (4): 889–902.

Morris, Stephanie A., Songjoon Baek, Myong-Hee Sung, Sam John, Malgorzata Wiench, Thomas A. Johnson, R. Louis Schiltz, and Gordon L. Hager. 2014. “Overlapping Chromatin-Remodeling Systems Collaborate Genome Wide at Dynamic Chromatin Transitions.” *Nature Structural & Molecular Biology* 21 (1): 73–81.

Moyle-Heyrman, Georgette, Hannah S. Tims, and Jonathan Widom. 2011. “Structural Constraints in Collaborative Competition of Transcription Factors against the Nucleosome.” *Journal of Molecular Biology* 412 (4): 634–46.

Muraoka, Naoto, Hiroyuki Yamakawa, Kazutaka Miyamoto, Taketaro Sadahiro, Tomohiko Umei, Mari Isomi, Hanae Nakashima, et al. 2014. "MiR-133 Promotes Cardiac Reprogramming by Directly Repressing Snai1 and Silencing Fibroblast Signatures." *The EMBO Journal* 33 (14): 1565–81.

Ng, Alex H. M., Parastoo Khoshakhlagh, Jesus Eduardo Rojo Arias, Giovanni Pasquini, Kai Wang, Anka Swiersy, Seth L. Shipman, et al. 2021. "A Comprehensive Library of Human Transcription Factors for Cell Fate Engineering." *Nature Biotechnology* 39 (4): 510–19.

Oudet, P., M. Gross-Bellard, and P. Chambon. 1975. "Electron Microscopic and Biochemical Evidence That Chromatin Structure Is a Repeating Unit." *Cell* 4 (4): 281–300.

Pagliuca, Felicia W., Jeffrey R. Millman, Mads Gürtler, Michael Segel, Alana Van Dervort, Jennifer Hyoje Ryu, Quinn P. Peterson, Dale Greiner, and Douglas A. Melton. 2014. "Generation of Functional Human Pancreatic β Cells In Vitro." *Cell* 159 (2): 428–39.

Partridge, E. Christopher, Surya B. Chhetri, Jeremy W. Prokop, Ryne C. Ramaker, Camden S. Jansen, Say-Tar Goh, Mark Mackiewicz, et al. 2020. "Occupancy Maps of 208 Chromatin-Associated Proteins in One Human Cell Type." *Nature* 583 (7818): 720–28.

Parviz, Fereshteh, Christine Matullo, Wendy D. Garrison, Laura Savatski, John W. Adamson, Gang Ning, Klaus H. Kaestner, Jennifer M. Rossi, Kenneth S. Zaret, and Stephen A. Duncan. 2003. "Hepatocyte Nuclear Factor 4α Controls the Development of a Hepatic Epithelium and Liver Morphogenesis." *Nature Genetics* 34 (3): 292–96.

Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression." *Nature Methods* 14 (4): 417–19.

Polach, K. J., and J. Widom. 1995. "Mechanism of Protein Access to Specific DNA Sequences in Chromatin: A Dynamic Equilibrium Model for Gene Regulation." *Journal of Molecular Biology* 254 (2): 130–49.

———. 1996. "A Model for the Cooperative Binding of Eukaryotic Regulatory Proteins to Nucleosomal Target Sites." *Journal of Molecular Biology* 258 (5): 800–812.

Posey, Ammon E., Alex S. Holehouse, and Rohit V. Pappu. 2018. "Chapter One - Phase Separation of Intrinsically Disordered Proteins." In *Methods in Enzymology*, edited by Elizabeth Rhoades, 611:1–30. Academic Press.

Qian, Li, Yu Huang, C. Ian Spencer, Amy Foley, Vasanth Vedantham, Lei Liu, Simon J. Conway, Ji-Dong Fu, and Deepak Srivastava. 2012. "In Vivo Reprogramming of Murine Cardiac Fibroblasts into Induced Cardiomyocytes." *Nature* 485 (7400): 593–98.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42.

Ramachandran, Srinivas, and Steven Henikoff. 2016. “Transcriptional Regulators Compete with Nucleosomes Post-Replication.” *Cell* 165 (3): 580–92.

Ramakrishnan, V., J. T. Finch, V. Graziano, P. L. Lee, and R. M. Sweet. 1993. “Crystal Structure of Globular Domain of Histone H5 and Its Implications for Nucleosome Binding.” *Nature* 362 (6417): 219–23.

Ramírez, Fidel, Devon P. Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S. Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. 2016. “deepTools2: A next Generation Web Server for Deep-Sequencing Data Analysis.” *Nucleic Acids Research* 44 (W1): W160–65.

Ramsköld, Daniel, Eric T. Wang, Christopher B. Burge, and Rickard Sandberg. 2009. “An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data.” *PLoS Computational Biology* 5 (12): e1000598.

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.

Robson, M. K., J. M. Anderson, O. M. Garson, J. P. Matthews, and T. F. Sandeman. 1981. “Constitutive Heterochromatin (C-Banding) Studies in Patients with Testicular Malignancies.” *Cancer Genetics and Cytogenetics* 4 (4): 319–23.

Rufibach, Laura E., Stephen A. Duncan, Michele Battle, and Samir S. Deeb. 2006. “Transcriptional Regulation of the Human Hepatic Lipase (LIPC) Gene Promoter.” *Journal of Lipid Research* 47 (7): 1463–77.

Ryan Corces, M., 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.

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.

Schultz, J. 1936. “Variegation in Drosophila and the Inert Chromosome Regions.” *Proceedings of the National Academy of Sciences of the United States of America* 22 (1): 27–33.

Sekiya, Sayaka, and Atsushi Suzuki. 2011. “Direct Conversion of Mouse Fibroblasts to Hepatocyte-like Cells by Defined Factors.” *Nature* 475 (7356): 390–93.

Sherwood, Richard I., Tatsunori Hashimoto, Charles W. O’Donnell, Sophia Lewis, Amira A. Barkal, John Peter van Hoff, Vivek Karun, Tommi Jaakkola, and David K. Gifford. 2014. “Discovery of Directional and Nondirectional Pioneer Transcription Factors by Modeling DNase Profile Magnitude and Shape.” *Nature Biotechnology* 32 (2): 171–78.

Shim, E. Y., C. Woodcock, and K. S. Zaret. 1998. “Nucleosome Positioning by the Winged Helix Transcription Factor HNF3.” *Genes & Development* 12 (1): 5–10.

Song, Kunhua, Young-Jae Nam, Xiang Luo, Xiaoxia Qi, Wei Tan, Guo N. Huang, Asha Acharya, et al. 2012. “Heart Repair by Reprogramming Non-Myocytes with Cardiac Transcription Factors.” *Nature* 485 (7400): 599–604.

Song, Lingyun, and Gregory E. Crawford. 2010. “DNase-Seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells.” *Cold Spring Harbor Protocols* 2010 (2): db.prot5384.

Soufi, Abdenour, Greg Donahue, and Kenneth S. Zaret. 2012. “Facilitators and Impediments of the Pluripotency Reprogramming Factors’ Initial Engagement with the Genome.” *Cell* 151 (5): 994–1004.

Soufi, Abdenour, Meilin Fernandez Garcia, Artur Jaroszewicz, Nebiyu Osman, Matteo Pellegrini, and Kenneth S. Zaret. 2015. “Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming.” *Cell* 161 (3): 555–68.

Stark, Rory, Gordon Brown, and Others. 2011. “DiffBind: Differential Binding Analysis of ChIP-Seq Peak Data.” *R Package Version* 100 (4.3). https://bioconductor.statistik.tu-dortmund.de/packages/2.13/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf.

Su, Zhida, Wenze Niu, Meng-Lu Liu, Yuhua Zou, and Chun-Li Zhang. 2014. “In Vivo Conversion of Astrocytes to Neurons in the Injured Adult Spinal Cord.” *Nature Communications* 5 (February): 3338.

Swinstead, Erin E., Tina B. Miranda, Ville Paakinaho, Songjoon Baek, Ido Goldstein, Mary Hawkins, Tatiana S. Karpova, et al. 2016. “Steroid Receptors Reprogram FoxA1 Occupancy through Dynamic Chromatin Transitions.” *Cell* 165 (3): 593–605.

Takahashi, Kazutoshi, and Shinya Yamanaka. 2006. “Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors.” *Cell* 126 (4): 663–76.

Tapscott, S. J., R. L. Davis, M. J. Thayer, P. F. Cheng, H. Weintraub, and A. B. Lassar. 1988. “MyoD1: A Nuclear Phosphoprotein Requiring a Myc Homology Region to Convert Fibroblasts to Myoblasts.” *Science* 242 (4877): 405–11.

Thorel, Fabrizio, Virginie Népote, Isabelle Avril, Kenji Kohno, Renaud Desgraz, Simona Chera, and Pedro L. Herrera. 2010. “Conversion of Adult Pancreatic Alpha-Cells to Beta-Cells after Extreme Beta-Cell Loss.” *Nature* 464 (7292): 1149–54.

Treutlein, Barbara, Qian Yi Lee, J. Gray Camp, Moritz Mall, Winston Koh, Seyed Ali Mohammad Shariati, Sopheak Sim, et al. 2016. “Dissecting Direct Reprogramming from Fibroblast to Neuron Using Single-Cell RNA-Seq.” *Nature* 534 (7607): 391–95.

Uhlén, Mathias, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, et al. 2015. "Proteomics. Tissue-Based Map of the Human Proteome." *Science* 347 (6220): 1260419.

Vaseghi, Haley Ruth, Chaoying Yin, Yang Zhou, Li Wang, Jiandong Liu, and Li Qian. 2016. "Generation of an Inducible Fibroblast Cell Line for Studying Direct Cardiac Reprogramming." *Genesis* 54 (7): 398–406.

Velazco-Cruz, Leonardo, Jiwon Song, Kristina G. Maxwell, Madeleine M. Goedegebuure, Punn Augsornworawat, Nathaniel J. Hogrebe, and Jeffrey R. Millman. 2018. "Acquisition of Dynamic Function in Human Stem Cell-Derived β Cells." *Stem Cell Reports*, December. https://doi.org/10.1016/j.stemcr.2018.12.012.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, et al. 2001. "The Sequence of the Human Genome." *Science* 291 (5507): 1304–51.

Vettese-Dadey, M., P. Walter, H. Chen, L. J. Juan, and J. L. Workman. 1994. "Role of the Histone Amino Termini in Facilitated Binding of a Transcription Factor, GAL4-AH, to Nucleosome Cores." *Molecular and Cellular Biology* 14 (2): 970–81.

Vierbuchen, Thomas, Austin Ostermeier, Zhiping P. Pang, Yuko Kokubu, Thomas C. Südhof, and Marius Wernig. 2010. "Direct Conversion of Fibroblasts to Functional Neurons by Defined Factors." *Nature* 463 (7284): 1035–41.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. 2020. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods* 17 (3): 261–72.

Wang, Li, Ziqing Liu, Chaoying Yin, Huda Asfour, Olivia Chen, Yanzhen Li, Nenad Bursac, Jiandong Liu, and Li Qian. 2015. "Stoichiometry of Gata4, Mef2c, and Tbx5 Influences the Efficiency and Quality of Induced Cardiac Myocyte Reprogramming." *Circulation Research* 116 (2): 237–44.

Wapinski, Orly L., Thomas Vierbuchen, Kun Qu, Qian Yi Lee, Soham Chanda, Daniel R. Fuentes, Paul G. Giresi, et al. 2013. "Hierarchical Mechanisms for Direct Reprogramming of Fibroblasts to Neurons." *Cell* 155 (3): 621–35.

Wilmut, I., A. E. Schnieke, J. McWhir, A. J. Kind, and K. H. Campbell. 1997. "Viable Offspring Derived from Fetal and Adult Mammalian Cells." *Nature* 385 (6619): 810–13.

Woodcock, C. L., J. P. Safer, and J. E. Stanchfield. 1976. "Structural Repeating Units in Chromatin. I. Evidence for Their General Occurrence." *Experimental Cell Research* 97 (January): 101–10.

Workman, J. L., T. J. Schuetz, and R. E. Kingston. 1991. "Facilitated Binding of GAL4 and Heat Shock Factor to Nucleosomal Templates: Differential Function of DNA-Binding Domains." *Genes*. http://genesdev.cshlp.org/content/5/7/1285.short.

Yan, Chao, Hengye Chen, and Lu Bai. 2018. “Systematic Study of Nucleosome-Displacing Factors in Budding Yeast.” *Molecular Cell* 71 (2): 294–305.e4.

Yoo, Andrew S., Alfred X. Sun, Li Li, Aleksandr Shcheglovitov, Thomas Portmann, Yulong Li, Chris Lee-Messer, Ricardo E. Dolmetsch, Richard W. Tsien, and Gerald R. Crabtree. 2011. “MicroRNA-Mediated Conversion of Human Fibroblasts to Neurons.” *Nature* 476 (7359): 228–31.

Yu, Xinyang, and Michael J. Buck. 2019. “Defining TP53 Pioneering Capabilities with Competitive Nucleosome Binding Assays.” *Genome Research* 29 (1): 107–15.

Zaret, Kenneth S., and Jason S. Carroll. 2011. “Pioneer Transcription Factors: Establishing Competence for Gene Expression.” *Genes & Development* 25 (21): 2227–41.

Zaret, Kenneth S., and Susan E. Mango. 2016. “Pioneer Transcription Factors, Chromatin Dynamics, and Cell Fate Control.” *Current Opinion in Genetics & Development* 37 (April): 76–81.

Zhang, Jing, Donghoon Lee, Vineet Dhiman, Peng Jiang, Jie Xu, Patrick McGillivray, Hongbo Yang, et al. 2020. “An Integrative ENCODE Resource for Cancer Genomics.” *Nature Communications* 11 (1): 3696.

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): 1–9.

Zhao, Yuanbiao, Pilar Londono, Yingqiong Cao, Emily J. Sharpe, Catherine Proenza, Rebecca O’Rourke, Kenneth L. Jones, et al. 2015. “High-Efficiency Reprogramming of Fibroblasts into Cardiomyocytes Requires Suppression of pro-Fibrotic Signalling.” *Nature Communications* 6 (September): 8243.

Zhu, Fangjie, Lucas Farnung, Eevi Kaasinen, Biswajyoti Sahu, Yimeng Yin, Bei Wei, Svetlana O. Dodonova, et al. 2018. “The Interaction Landscape between Transcription Factors and the Nucleosome.” *Nature* 562 (7725): 76–81.

Zhu, Jiang, Fuhong He, Shuhui Song, Jing Wang, and Jun Yu. 2008. “How Many Human Genes Can Be Defined as Housekeeping with Current Expression Data?” *BMC Genomics* 9 (April): 172.