The Pennsylvania State University

The Graduate School

SYSTEMATIC DISSECTION OF SEQUENCE FEATURES AFFECTING THE BINDING SPECIFICITY OF A PIONEER FACTOR

A Dissertation in

Molecular, Cellular and Integrated Biosciences

by

Cheng Xu

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2024

The dissertation of Cheng Xu was reviewed and approved by the following:

Lu Bai
Professor of Biochemistry and Molecular Biology
Professor of Physics
Dissertation Advisor
Chair of Committee

Shaun Mahony
Associate Professor of Biochemistry and Molecular Biology

Xiaojun Lian
Associate Professor of Biomedical Engineering

Robert F. Paulson
Professor of Veterinary and Biomedical Sciences

Melissa Rolls
Paul Berg Professor of Biochemistry and Molecular Biology
Chair of the Graduate Program

ABSTRACT

Sequence-specific transcription factors (TFs), which recognize their cognate DNA motifs, are central players in regulation of gene expression. In higher eukaryotes, TFs only bind to a small proportion of their motifs across the genome, partially because of the inhibition of nucleosomes. Strong binding selectivity is also observed for pioneer factors (PFs) despite their ability to bind to nucleosomal DNA, and the underlying mechanism is not well understood. Here, we designed a high-throughput assay named Chromatin Immunoprecipitation with Integrated Synthetic Oligonucleotides (ChIP-ISO) to systematically dissect local sequence features affecting the binding specificity of a classic PF, FoxA1, in A549 human lung carcinoma cells. This method involves integrating thousands of synthetic sequences containing FoxA1 motifs into a fixed genomic locus, followed by FoxA1 chromatin immunoprecipitation (ChIP) and next-generation amplicon sequencing.

We found that within the same sequence background, FOXA1 binding is strongly affected by its motif strength, clustering of motifs, and co-binding TFs including AP-1 and CEBPB. AP-1 is particularly important for enhancing FOXA1 binding, which is further illustrated by inhibition of AP-1 binding and genome-wide studies. Comparison among different cell lines and RNA-seq analysis reveal that AP-1 contributes to the cell-type-specific binding and functions of FOXA1. *In vivo* and *In vitro* studies further confirmed the interdependency and cooperativity between FOXA1 and AP-1 binding, although FOXA1 binding to naked DNA depends more on its core motifs. Finally, by moving sequences originated from different genomic loci to the same chromatin background and measuring FOXA1 binding, we showed that FOXA1's binding specificity is more determined by the local sequence than chromatin background, including H3K9me3 or H3K27me3-marked heterochromatin. Our conclusions are consistent with a convolutional neural network (CNN) analysis of FOXA1 ChIP-seq data.

In summary, our study provides insights of the genetic rules underlying PF binding specificity and reveals a potential mechanism for regulating its binding events during cell differentiation. Our study also establishes an experimental framework for understanding TF binding specificity and cis-regulatory logic.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

My advisor, Dr. Lu Bai, is definitely the most essential part of my memorable Ph.D experience. Her enthusiasm about science, innovative scientific thinking, and critical and rigorous altitude guides me through the way to become an independent researcher. I'm grateful for her patience, giving me freedom to follow and test my own scientific thoughts, and supporting me in all the aspects. This experience of working with her certainly shapes my life, and inspires me to keep pursuing my scientific dream.

I would like to thank my committee members, Dr. Shaun Mahony, Dr. Xiaojun Lian and Dr. Robert Paulson, as well as former committee members, Dr. David Gilmour and Dr. Yanming Wang for their commitment in helping me fulfill my dissertation. Their insightful suggestions and expertise in different fields really pushed the project to where I cannot reach alone.

I want to express my special gratefulness to Dr. Mahony, who acts both as my committee member and collaborator. I really enjoy the productive communication with him on my project and also want to thank him for always being responsive and supportive to me.

I also want to give special thanks to Dr. Wang for training and getting me started on mammalian tissue culture, as well as sharing their equipment and reagents.

I would like to thank Holly Godin, who work side-by-side with me on this project. We went through both encouraging moments and hard time. She is always dedicated, trustworthy, optimistic and constructive. I'm worried from time to time that I would never find such a great collaborator in my future career.

I would like to thank Jenna Johnson, the undergraduate researcher who worked with me. She is really talented and reliable in carrying out independent research, which contributes significantly to this work.

I would like to thank all the other collaborators, Jianyu Yang from Mahony Lab and Erik Leith from Dr. Song Tan's lab. Their contributions are inseparable parts of this project.

I would like to thank Dr. Cheryl Keller and Dr. Craig Praul for helping me with next-generation sequencing.

I would like to thank all the Bai Lab members and alumni. We are both colleagues and good friends. I will always remember the days we spent together, working hard for scientific discoveries, helping each other out, and sharing happiness and glory.

I would like to thank all the members of Center for Eukaryotic Gene Regulation for creating an excellent research environment, and providing invaluable suggestions for my research.

I would like to thank program chair Dr. Melissa Rolls and program coordinator Freya Heryla for always being responsive and supportive to me as a graduate student.

I would like to thank my parents. Studying abroad for a PhD degree means separating from them for many years. I'm deeply grateful for them unconditionally supporting my choice to pursue my dream as a scientist.

I would like to thank my girlfriend, Sangshan Tian. She is indeed my "Sunshine" in life. Meeting and having her as my life partner is the luckiest thing I have ever had.

I would like to thank my uncle's family for taking care of me over the years as my closest relatives in the United States.

I would like to thank Dr. Manyu Du, Dr. Jun Wang, Dr. Haoyang Jiang, Dr. Feiyue Lu, Dr. Lai Shi and Dr. Jian Sun, who are both my friends and colleagues in the close field. I greatly appreciate their guidance in my research, career and life.

I would like to thank my cohort and all other friends. I will always cherish the shiny memory we created together.

This dissertation work is supported by the National Institutes of Health (R35 GM139654 to Dr. Lu Bai) and the Graduate Research Innovation fund from the Huck Institute of Life Sciences (to Cheng Xu). The findings and conclusions do not necessarily reflect the view of the funding agency.

ABBREVIATIONS

Basic helix-loop-helix	bHLH
Basic leucine zipper	bZIP
Blasticidin resistance gene	*BlaR*
C2H2-zinc finger	C2H2-ZF
CCND1 enhancer	*CCND1e*
ChIP with microarray	ChIP-chip
ChIP with next-generation sequencing	ChIP-seq
Chromatin immunoprecipitation	ChIP
Chromatin Immunoprecipitation with Integrated Synthetic Oligonucleotides	ChIP-ISO
Cleavage under targets and release using nuclease	CUT&RUN
Cleavage under targets and tagmentation	CUT&Tag
Co-immunoprecipitation	co-IP
Consecutive affinity-purification systematic evolution of ligands by exponential enrichment	CAP-SELEX
Consensus	con
Convolutional neural network	CNN
Cross-linking immunoprecipitation-MS	xIP-MS
Cross-linking mass spectrometry	CLMS
Cryogenic electron microscopy	cryoEM
DNA-binding domain	DBD
Downstream promoter element	DPE

Doxycycline	Dox
Electrophoretic mobility shift assay	EMSA
Electrophoretic mobility shift assay followed by sequencing	EMSA-seq
Embryonic Stem Cell	ESC
Encyclopedia of DNA Elements	ENCODE
Escherichia coli	*E. coli*
Extradenticle	Exd
Fibroblasts	Fib
Fluorescence activated cell sorting	FACS
Gene ontology	GO
General transcription factors	GTF
Glucocorticoid receptor	GR
Grany head	Grh
Homology-directed repair	HDR
Induced pluripotent stem cell	iPSC
Initiator	Inr
Interferon-β	IFN-β
Lysine 270 of FOXA1	FOXA1-K270
Micrococcal nuclease	MNase
Motif ten element	MTE
Original	ori
Pioneer factor	PF
Position weight matrix	PWM
Post-translational modifications	PTM

Precursor messenger RNA	pre-mRNA
Preinitiation complex	PIC
Promoterless puromycin resistance gene	*PuroR*
Protective PFs	PPF
Protein binding microarray	PBM
RNA polymerase II	Pol II
Steroid receptors	SR
Super PFs	SPF
TATA binding protein	TBP
Transcription factor	TF
Transcription start site	TSS
Transcripts per million	TPM
Transposable elements	TE
Transposase-accessible chromatin using sequencing	ATAC
Wild-type	WT

Chapter 1

Introduction

Work presented in this chapter includes a part of the published manuscript "Kleinschmidt, H., Xu, C., & Bai, L. (2023). Using Synthetic DNA Libraries to Investigate Chromatin and Gene Regulation. Chromosoma, 1-23.", and a part of the manuscript under review "Stoeber, S., Godin, H., Xu, C., & Bai, L. (2024). Pioneer Factors: Nature or Nurture.".

1.1 Regulation of Gene Expression and Transcription Factors

1.1.1 Storage and flow of genetic information

Throughout the course of human civilization, there has been a persistent curiosity regarding how the traits of life are determined and inherited. However, it remained mysterious until the 19th century, when Gregor Mendel established the laws of Mendelian inheritance with his pea plant hybridization experiments[1]. The inheritable substance that carried genetic information in his hypothesis (called "elementen") was later found to be part of chromosomes by Thomas Hunt Morgan[2] in 1910, and determined to be DNA by Oswald Avery and his colleagues in 1944[3]. In 1953, the double-helix model of DNA structure was reported by James Watson and Francis Crick[4], elucidating the molecular basis of this fundamental building blocks of life. These significant discoveries revealed the elegant way in which genetic information is stored in living organisms.

Building on these findings, subsequent research has centered on investigating how genetic information is transferred from DNA to proteins, which are major participants of virtually

all biological processes. In 1958, Francis Crick proposed the central dogma of molecular biology, which describes the flow of genetic information inside a biological system[5]. In this theory, sequence information stored in DNA can be transferred to RNA through transcription, and further passed to proteins via translation. Based on the theory, the genetic code of this process was then deciphered by Marshall Nirenberg and Har Gobind Khorana[6-8], establishing a direct connection between DNA sequence and the sequence of amino acids in proteins. These fundamental breakthroughs not only enabled unraveling of gene functions, but also empowered scientists to employ and manipulate genes for research, medical and engineering applications.

1.1.2 First example of gene regulation

Transmission of genetic information from DNA to RNA and protein, i.e., gene expression, is also a tightly regulated process in living organisms. In fact, with differential regulation, the same genome can be decoded in various ways, leading to enormously diverse functions. Sophisticated regulation of gene expression patterns underlies almost all cellular and biological processes including cellular structure, differentiation, development, metabolism and responses to the environment, while dysregulated gene expression gives rise to diseases.

The first well-understood gene regulation mechanism stems from François Jacob and Jacques Monod's seminal study on the *lac* operon in *Escherichia coli* (*E. coli*)[9]. The *lac* operon consists of a cluster of genes that share the same promoter and are involved in lactose metabolism. Jacob and Monod found that in the absence of lactose, a protein named *lac* repressor binds to the operator region upstream of the *lac* genes, blocking their transcription. When lactose is present, the *lac* repressor gets bound by allolactose (converted from lactose), goes through a conformational change, and loses the ability to bind to the operator. This enables the transcription of the *lac* genes to occur. This prototypical model emphasized the importance of investigating

mechanisms of gene regulation for comprehending the function of the genome, as stated in the original paper:

> "The discovery of regulator and operator genes, and of repressive regulation of the activity of structural genes, reveals that the genome contains not only a series of blue-prints, but a co-ordinated program of protein synthesis and the means of controlling its execution."

1.1.3 Gene regulation in eukaryotes

The regulation of eukaryotic gene expression is underscored by more intricate and diverse mechanisms. Importantly, the regulation process happens at multiple levels including transcription, RNA splicing[10], RNA degradation[11], translation[12] and post-translational modification of protein[13]. Among them, transcription, serving as the initial stage of gene expression, plays a primary role.

Study of eukaryotic transcription mechanism started with the discovery of three chromatographically distinct RNA polymerases, RNA polymerase I (Pol I), Pol II and Pol III, by Robert Roeder in 1969[14]. Pol II was later shown to be responsible for producing precursor messenger RNA (pre-mRNA)[15], thus is required for transcribing all the protein-coding genes. Due to its essential role, eukaryotic transcription is commonly referred to as transcription mediated by Pol II. Pol II Transcription is a multistep process. An intact transcription cycle is comprised of chromatin opening, preinitiation complex (PIC) assembly, initiation, promoter-proximal pausing, elongation, termination and recycling[16] (**Figure 1-1**). All of these steps are under rigorous regulation[17-20]. Notably, the formation of PIC usually acts as a major rate-limiting step in transcription activation[17], and has been extensively studied.

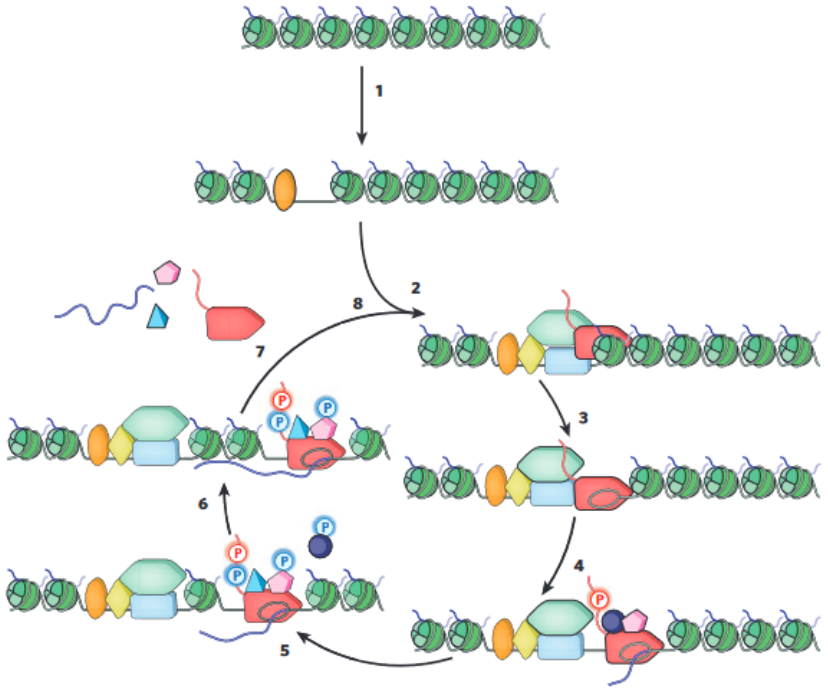


Figure 1-1: An intact transcription cycle.

Picture from Ref[16].

PIC is assembled by Pol II and a set of general transcription factors (GTFs), which include TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH, at the core promoter[21]. The core promoter is a short sequence that covers a region from ~50 bp upstream to ~50 bp downstream of the transcription start site (TSS). It contains motifs like TATA-box, initiator (Inr), downstream promoter element (DPE) and motif ten element (MTE), whose positions are fixed with respect to the TSS. One thing to notice is that none of these motifs exists universally in all core promoters, and the compositions of different core promoters are highly diversified[22]. On a TATA-containing promoter, with the TATA binding protein (TBP) subunit of TFIID binding to the TATA-box, TFIIA and TFIIB are sequentially incorporated into the TBP-promoter complex, which then recruits Pol II with the assistance of TFIIF. Finally, TFIIE and TFIIH are integrated, forming a

functional PIC that can induce basal transcription[17] (**Figure 1-2**). On TATA-less promoters, PIC assembly required the entire TFIID complex rather than solely the function of the TBP subunit[23].

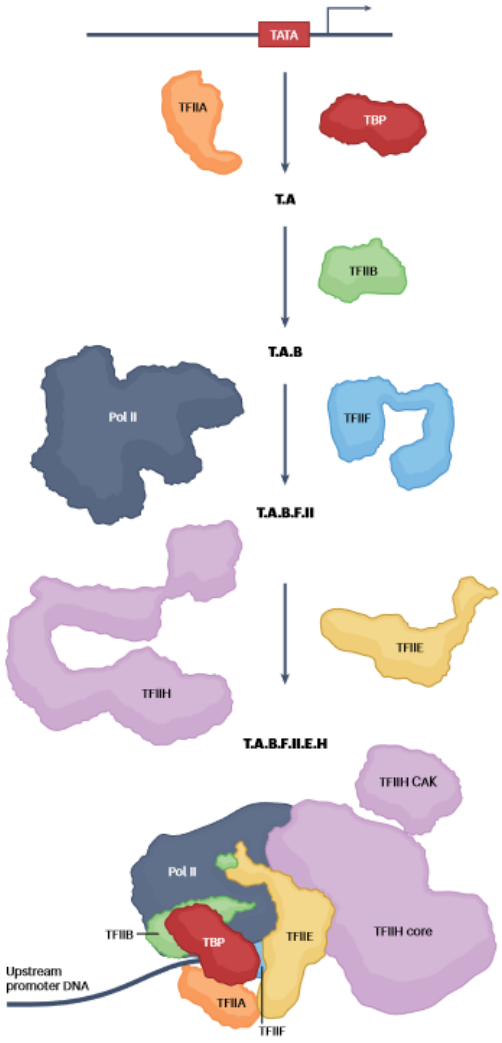


Figure 1-2: Preinitiation complex (PIC) assembly.

Picture from Ref[17].

The assembly of PIC generates basal transcriptional activity. The full activation of transcription, in turn, requires sequence-specific transcription factors (TFs), Mediator and coactivators[23] (**Figure 1-3**). Sequence-specific TFs, also known as gene-specific transcriptional activators, bind to *cis*-regulatory elements (CRE) that are proximal (promoters) or distal

(enhancers) to the TSS[24], and determine the spatial and temporal patterns of gene expression[17]. Mediator is a multi-subunit complex that connects between the gene-specific regulatory signals from sequence-specific TFs and PIC[25]. It works together with other coactivators to ensure full activation and precise regulation of transcription[17].

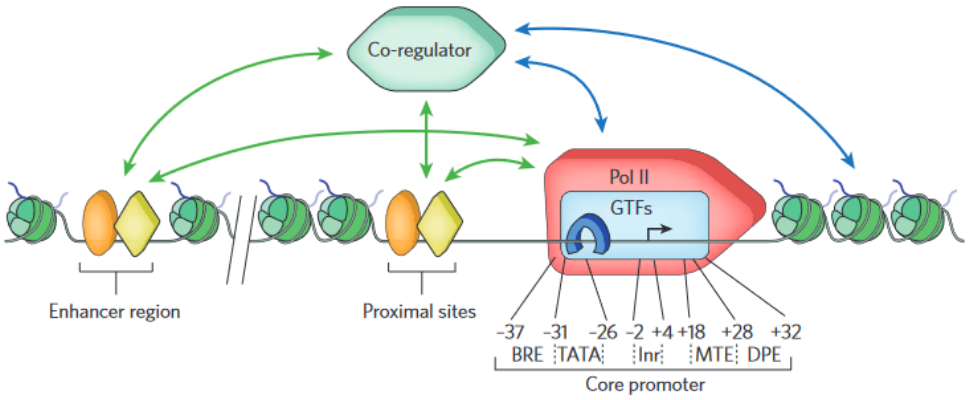


Figure 1-3: Mechanism of transcription initiation.

Picture from Ref[16].

1.1.4 Sequence-specific transcription factors (TFs)

Sequence-specific TFs, as their name indicate, recognize specific DNA sequences to regulate transcription[26]. While the term TF can be used to describe both GTFs and sequence-specific TFs, it often just refers to sequence-specific TFs by researchers. They are key players in gene expression regulation, as they decide what genes to be expressed, when, where and the extent to which they should be expressed, which establishes specific gene expression patterns and regulatory networks in specific cell types. Therefore, TFs usually carry essential functions in development, physiological processes and cell reprogramming[27-29], and mutations or dysregulation of TFs lead to diseases[30].

A prototypical TF is usually composed of DNA-binding domain(s) and effector domain(s). The DNA-binding domain (DBD) is responsible for binding DNA in a sequence-specific manner, and frequently used for the classification of TFs. The effector domain generates activating or inhibitory effects on transcription, through processes including steric hindrance and interaction with coactivators or corepressors. It can also regulate the activity of the TF in response to signals and carry enzymatic activities[26,31]. Since the early discovery of TFs including glucocorticoid receptor (GR[32]), SP1[33], HSF[34], USF/MLTF[35,36] and Gal4[37] in the 1980s, a huge collection of eukaryotic TFs has been characterized so far. A recent comprehensive review on human TFs[26] identified in total 1,639 known or likely human TFs. While early studies mainly identified new TFs based on their *in vitro* DNA binding and transcriptional regulation activities, nowadays, discoveries of new TFs are largely done by testing sequence homology to known DBDs.

The 1,639 known or putative human TFs can be classified into a number of different families (**Figure 1-4**). The C2H2-zinc finger (C2H2-ZF) and Homeodomain family TFs constitute most of the human TFs. Besides these two families, bHLH, bZIP, Forkhead, nuclear hormone receptor, HMG/Sox and ETS families are the most common families of human TFs[26]. C2H2-ZF family TFs, which could contain 1 to 40 zinc finger domains binding DNA in a tandem array, is the largest TF family in the human genome[38]. Interesting, C2H2-ZF TFs also have the most diverse binding motifs, and many of the C2H2-ZF TFs containing KRAB domains exhibit signatures of diversifying selection[39] during evolution. It is proposed that these KRAB containing C2H2-ZF TFs specifically recognize transposable elements (TEs) and use the repressive effect of the KRAB domains to inhibit their activities. The rapid expansion of these C2H2-ZF TFs starts with Amniota, and is related to that placenta increases the possibility for the transmission of retrovirus[40,41]. In contrast, most of the remaining TFs from various families start to diverge at the base of Bilateria, which is around the time when cell-type diversity increases quickly[26]. Examples

of the most conserved TFs include the HOX TFs, which are responsible for determining specific regions along the head-tail axis during establishment of embryo body plans in metazoans[42].

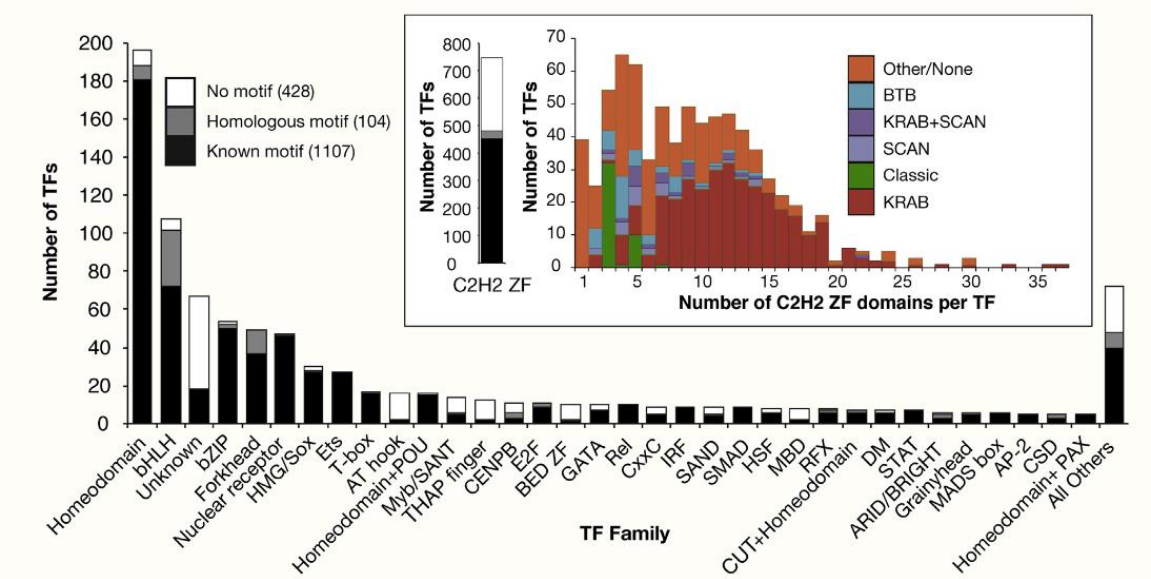


Figure 1-4: Families of human transcription factors.

Picture from Ref[26].

1.1.5 DNA-binding specificities of transcription factors

The DNA-binding specificity of a TF is one of the most important features that regulate the functions of the TF. The intrinsic DNA-binding preferences of a TF act as the primary layer of determinants for its targets in the genome. Typically, a TF is able to bind a collection of short and similar DNA sequences, which can be summarized as binding site motifs[43]. The most commonly used model for generating binding site motifs is the position weight matrix (PWM) model[44]. A PWM model describes the nucleotide preferences of the TF at each position of its motif, which can subsequently be represented by a sequence logo[45]. Currently, out of the 1,639 known or putative human TFs, 1,211 TFs have been assigned with a binding motif, either

measured by experiments, or inferred from homologs. Except C2H2-ZF TFs, most TFs from the same families or subfamilies recognize similar binding motifs[26] (**Figure 1-5**). The rich information about TF motifs from different species is stored in databases such as CIS-BP[46] and JASPAR[47], and can be analyzed by motif tools like MEME Suite[48].

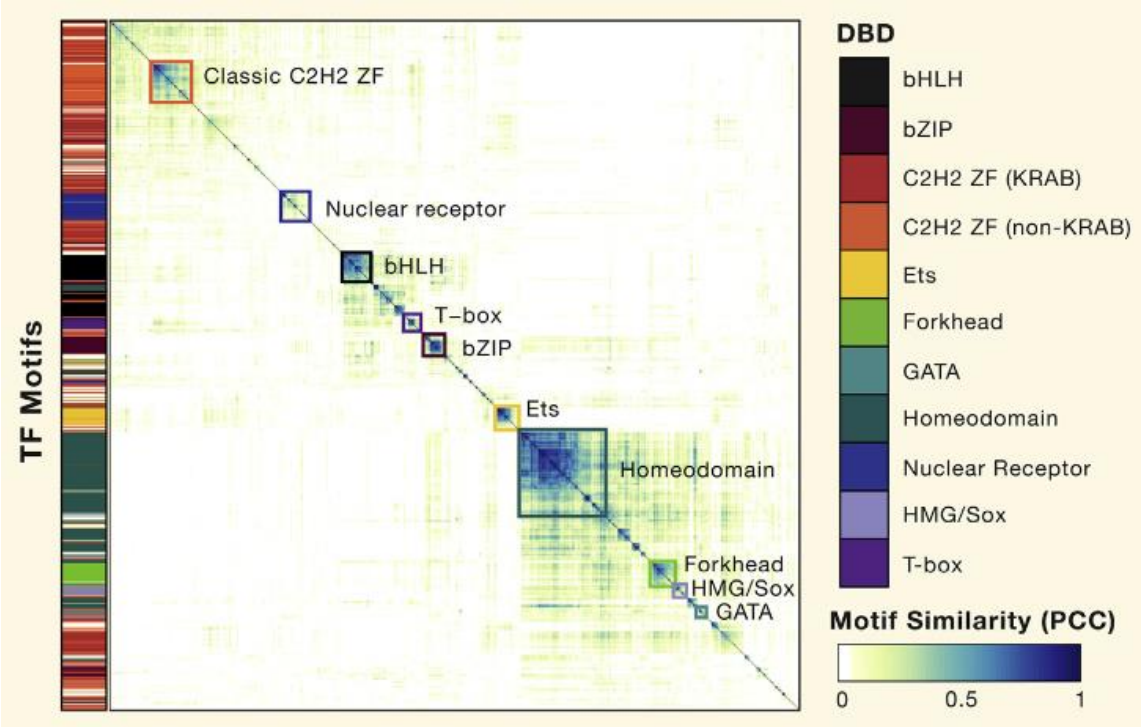


Figure 1-5: Motif similarity between human transcription factors.

Picture from Ref[26].

The binding motifs of TFs can be determined experimentally both *in vitro* and *in vivo*. Most of the existing motifs are generated by *in vitro* experiments that use purified proteins and DNA libraries[26], which measures intrinsic TF-DNA binding affinities in an unbiased manner. Widely used *in vitro* methods include systematic evolution of ligands through exponential enrichment (SELEX and its high-throughput version HT-SELEX)[49], protein binding microarray (PBM)[50] and bacterial one-hybrid[51]. In a SELEX experiment, a purified TF is incubated with a DNA library consisting of random sequences. Sequences bound by the TF will be selected out

and used as the DNA library for the next round of selection. Such selection will be performed for multiple rounds and followed by sequencing to eventually derive the PWM for the TF. Besides these high-throughput methods, TF binding specificities can also be studied *in vitro* with low-throughput methods such as electrophoretic mobility shift assay (EMSA)[52]. The implementation of EMSA is based on that DNA bound by protein will have less electrophoretic mobility compared to free DNA, thus generating a shifted band on the gel.

Chromatin immunoprecipitation (ChIP) is the most widely used method for profiling TF binding *in vivo*[53]. It involves crosslinking of protein-DNA interaction by formaldehyde, chromatin fragmentation by sonication or micrococcal nuclease (MNase), and immunoprecipitation by an antibody against the protein of interest to enrich for bound DNA. ChIP can be coupled with microarray (ChIP-chip) or next-generation sequencing (ChIP-seq) to characterize genome-wide TF binding sites. To increase the resolution of ChIP-seq, an additional step involving trimming the fragmented ChIP DNA by exonuclease has been added to the protocol of ChIP-exo and ChIP-nexus[54]. Aside from ChIP-based methods, a few other methods involving tethering enzymes to the TFs to measure their target sites have also been developed, including DamID[55], cleavage under targets and release using nuclease (CUT&RUN)[56] and cleavage under targets and tagmentation (CUT&Tag)[57].

1.1.6 Transcription factors only bind a small proportion of their potential motifs in the genome.

An unsolved problem regarding the binding specificities of TFs is that the PWM models derived *in vitro* have very weak predictive power for TF binding sites *in vivo*. In fact, in higher eukaryotes, out of the many occurrences of potential motifs across the genome, only a very small fraction is bound by the TF. For example, analysis of ChIP-seq data in K562 cells from the

Encyclopedia of DNA Elements (ENCODE) project revealed that on average 99.8% of putative binding motifs in the genome are not bound by the respective TF[58]. Moreover, the binding patterns for a specific TF are highly variable across different cell types[59]. Understanding the mechanism underlying the binding selectivity and cell-type specificity of TFs is critical for characterization of their functions *in vivo*.

A few mechanisms beyond the core binding motif have been proposed to be responsible for the selective and cell-type specific binding of TFs (**Figure 1-6**).

DNA shape

DNA shape features, including inter-base pair features (e.g. helix twist), intra-base pair features (e.g. propeller twist) and minor groove width[60], may affect TF binding to similar sequences. For example, anterior Hox proteins forming heterodimers with Extradenticle (Exd) selectively bind sequences with a special narrow minor grove in their binding motifs, which is distinct from other Exd-Hox complexes. The effect of DNA shape can be teased apart from the effect of sequence by mutating the residue that only recognize DNA shape[61,62].

Cooperative binding with other TFs

TF cooperative binding is a common phenomenon in higher eukaryotes, which is critical for determining DNA-binding specificity as well as the function after binding[63]. The classic example of TF cooperativity, the interferon-β (IFN-β) enhanceosome, involves cooperative binding of eight TFs to a short enhancer sequence through DNA conformational changes, which leads to recruitment of coactivators[64]. TF cooperative binding can result from direct protein-protein interaction, or interaction-independent mechanisms. In brief, cooperative TFs could form

stable complexes, interact with the guidance of DNA, contact indirectly through changes in DNA structure, or collaborate to deplete nucleosomes[65]. A high-throughput screen of human TFs discovered a large number of TF cooperative binding pairs and revealed that cooperativity can happen promiscuously between TFs from different families[66]. The cooperative binding of TFs mediates important biological processes such as face and limb development[67].

Other features including DNA methylation[68], nucleosome occupancy and chromatin accessibility[69], histone modifications[70], 3D genome contact[71], and variations in local TF concentrations[72] may also contribute to this binding selectivity.

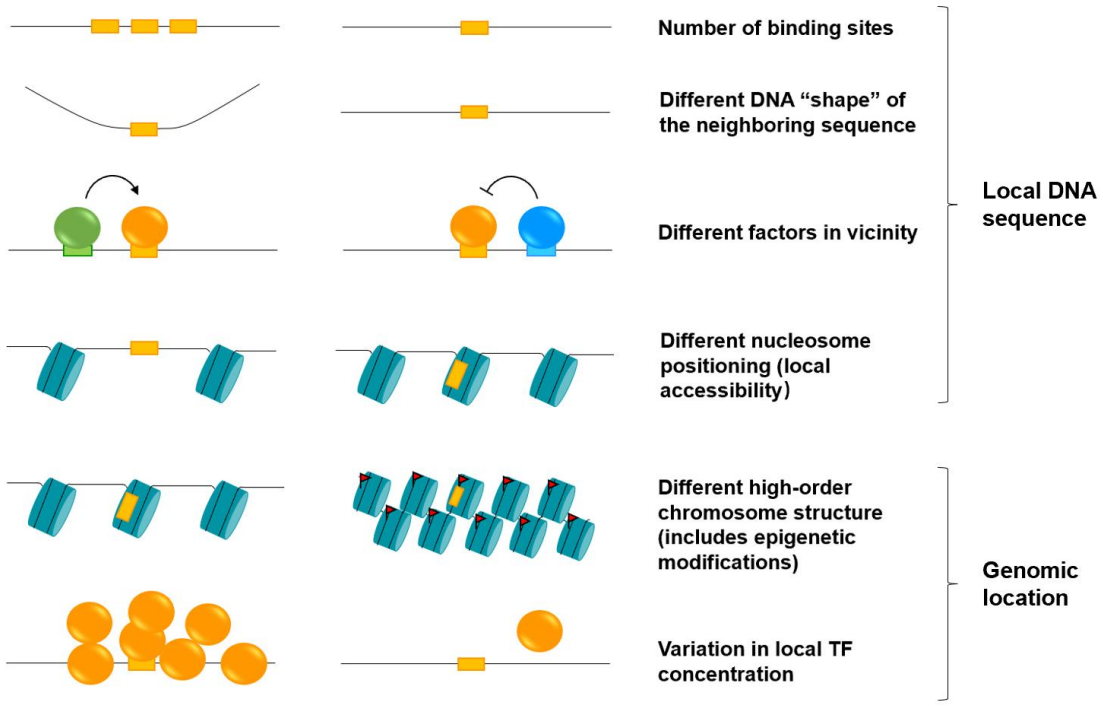


Figure 1- 6: Features that could potentially influence TF binding specificity.

1.2 Pioneer Transcription Factors

1.2.1 Pioneer factors

In eukaryotic cells, genomic DNA is packaged into nucleosomes. Nucleosomes act as barriers to many cellular events including transcription factor binding and transcription[73,74]. Local nucleosome occupancy serves as a major impediment to TF binding because a portion of the nucleosomal DNA surface is sterically occluded by the globular domains of histones, obstructing access to residues within the TF's recognition motif[75].

Due to the barrier nucleosomes create to transcription factor binding, accessing nucleosomal DNA and opening up the chromatin structure act as the primary steps for gene expression in eukaryotic organisms. A subset of transcription factors named pioneer factors (PFs) are able to overcome such barriers and bind nucleosomal DNA. Upon their binding, PFs can open up the local chromatin and enable other factors to access, endowing genes with the competence to be expressed[76,77] (**Figure 1-7**). Overall, PFs play essential roles in eukaryotic gene regulation. They are usually considered as master regulators of cellular activities in terms of their ability to shape transcriptional landscapes in certain cell types[76-78]. Therefore, studying PFs has significant meanings in understanding how gene expression and cellular activities are regulated.

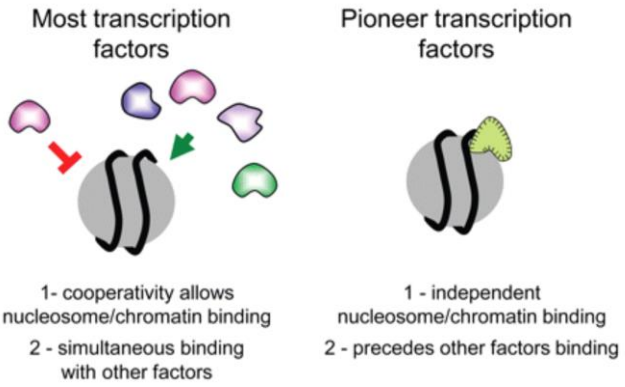


Figure 1-7: Comparison between pioneer factors and most other transcription factors.

Picture from Ref[76].

In mammalian systems, PFs have been studied extensively during development and cell reprogramming, as well as in cancer cells. One famous example demonstrating the role of PFs during cell reprogramming is the induced pluripotent stem cell (iPSC) technology. In this Nobel Prize-winning work, four transcription factors Oct3/4, Sox2, c-Myc and Klf4 were found to be sufficient to convert fibroblasts into iPSCs in mouse and human systems[29,79]. Subsequent studies showed that among the four factors, Oct3/4, Sox2 and Klf4 are PFs that promote cell reprogramming[80,81]. Considering their potential reprogramming activities, characterization of PFs from a large amount of transcription factors will assist in identifying cocktails of factors for cell reprogramming, which can be further applied to disease modeling, personalized medicine, and autologous therapies. The function of PFs has also been implicated in cancers, especially in hormone-dependent cancers. A few PFs including FoxA1, Pbx1 and AP2γ have been reported to promote cancer cell proliferation and metastasis[82-86]. It is possible to use these PFs as diagnostic biomarkers and drug targets for cancer therapies.

1.2.2 Mechanisms underlying nucleosome binding by pioneer factors.

Known pioneer factors have various types of DNA-binding domains[77]. Current evidence seems to suggest that different PFs use diverse mechanisms to access DNA embedded in nucleosomes (**Figure 1-8**). For example, FOXA1 and the H1 linker histone may use similar mechanisms to bind nucleosomes, as the DNA-binding domain of FOXA1 resembles linker histone structurally[87]. OCT4, SOX2 and KLF4 can recognize their partial motifs on the surface of nucleosomes[88]. One common feature for many PFs is the usage of short α-helix to bind DNA, so

that it will not collide with components of the nucleosome[89]. Some PFs have also been shown to have interactions with core histones, which contributes to their nucleosome targeting[90-92].

Another emerging mechanism underlying nucleosome binding is dissociation rate compensation. Budding yeast PFs Reb1 and Cbf1 can target nucleosomal DNA with similar affinities relative to naked DNA by utilizing a site exposure mechanism[93] in which the factors can bind to their nucleosome-embedded site following transient site exposure. Despite the reduced binding rate of Reb1 and Cbf1 to nucleosomes, these factors compensate for their reduced binding by reducing their dissociation rates[93]. The kinetic properties demonstrated by Reb1 and Cbf1 in vitro are also consistent with their ability to open chromatin *in vivo*[23,93]. Similar findings for a dissociation rate compensation mechanism were also recently reported for the Drosophila pioneer factor GAF[94] and SP1 and CTCF in U2OS cells[95], and FoxA1[96]. These studies demonstrate that PFs may be kinetically distinct from other sequence-specific TFs in their ability to compensate for reduced binding to chromatin by more stable interactions with their nucleosomal motifs.

On the other hand, PFs may also access nucleosomes by passive mechanisms. Nucleosomes are inherently dynamic and exist in states of fluctuation between being fully wrapped and partially unwrapped[97]. The fluctuation between wrapped and partially unwrapped states is sufficient to allow transcription factors that were once occluded by the nucleosome to bind their exposed motif. Additionally, if the factor is abundant enough and displays high specificity for its motif, then it can shift the equilibrium from a mostly wrapped state to that of an unwrapped state[98]. This model termed 'site-exposure'[99] is most prominent for factors with binding sites near the entry/exit sites of the nucleosomes. Studies have shown that the binding of passive transcription factors to nucleosome substrates decreases exponentially as the factor-binding site is moved further into the core of the nucleosome[98,99]. The contribution of nucleosome dynamics in vivo is still unclear. A recent study in yeast reports that a majority of the

nucleosomes in the yeast genome lack stable confirmations and that most nucleosomes have a portion of their DNA partially detached from the histone octamer[100]. The degree to which nucleosome dynamics facilitates the binding of both TFs and PFs to chromatin is likely influenced locally and context specifically.

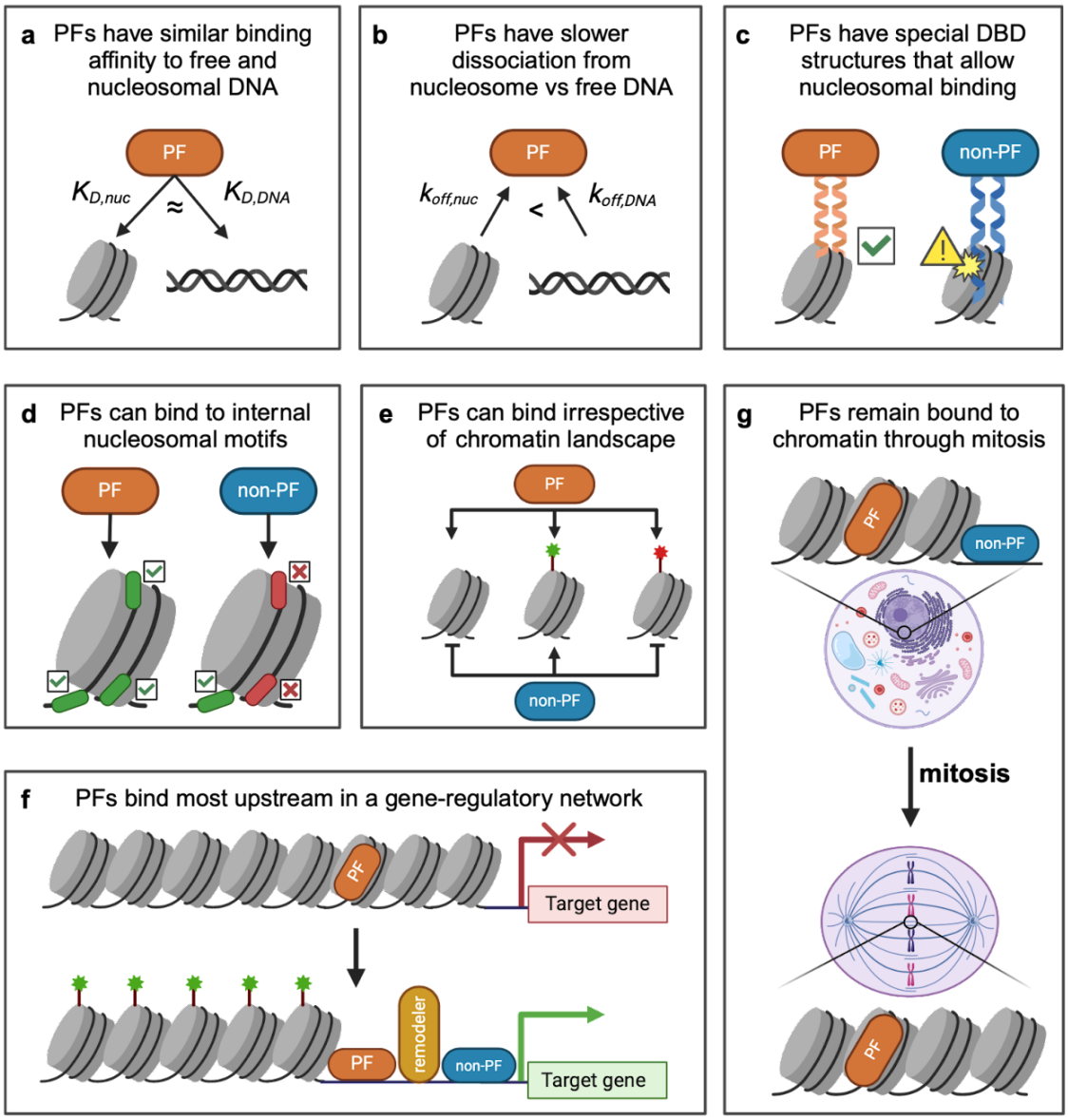


Figure 1- 8: Special properties of PFs that distinguish them from non-PFs.

Picture from Stoeber, S., Godin, H., Xu, C., & Bai, L. (2024). Pioneer Factors: Nature or Nurture. Under review in *Critical Reviews In Biochemistry & Molecular Biology*.

1.2.3 Pioneer factor motif position preferences within nucleosome

Several PFs prefer to bind to their motif when it is situated at a specific position or orientation within the underlying nucleosome. The combined rotational, translational, and orientational motif positioning preference of a PF on a nucleosome is called its "binding mode." Common translational binding modes include entry-exit, periodic, dyad, and dual gyre. These translational binding preferences may allow a PF to interact with or avoid steric clashes with specific histone residues or co-factors[101], or they may be constrained by the ability of a PF to bind a partial motif. Many PFs prefer to bind at the entry-exit site of the nucleosome where DNA may more easily unwrap from the nucleosome or transiently "breathe" to allow binding[102], including yeast PF Reb1[103,104]; human PFs TP53[105], OCT4[106,107], and GATA3[108]; and certain human basic helix-loop-helix (bHLH) and basic leucine zipper (bZIP) family members that bind >180° of the DNA surface[109]. Several PFs, like SOX family members, FOXA1, GATA3, and RFX5, prefer to bind at the nucleosomal dyad, where interference from histone residues or neighboring DNA is minimized[77,96,109,110 107]. Other PFs, like OCT4, have the ability to bind to motifs located at several positions across the nucleosomal surface[107,110,111]. Certain PFs, including OCT4 and T-box family factors, possess a unique "dual gyre" translational binding mode in which the PF interacts with both gyres of DNA on the nucleosome at the same time[109,111]. Dual gyre binding may be accomplished through the binding of each gyre by an individual TF domain or through dimerization of the factor, often to two individual motifs or partial motifs 80-bp in distance[109,111].

Translational positioning also affects rotational orientation. The majority of PFs, like SOX2, prefer rotational modes where their whole or partial motif is solvent-accessible versus modes where their motif is core-facing due to the physical occlusion of core-facing motifs by

histones[110-112]. Along this line, PFs that can bind across the nucleosomal surface often display periodic binding in which the PF preferentially binds to positions ~10bp apart, including human PFs EOMES, OCT4, and homeodomain family members, reflecting the exposure frequency of the minor or major groove of nucleosomal DNA[107,109]. Other PFs, like OCT4, do not exhibit preferred rotational modes, which could be due to their ability to recognize shorter, degenerate motifs that are more likely to be solvent accessible[111].

In many cases, PFs with asymmetrical motifs can bind to either orientation of their motif on nucleosomes, but certain PFs prefer to bind to a specific orientation. Motif orientation dictates the orientation of the PF relative to nearby nucleosomal residues, so certain motif orientations can lead to steric clashing and prevent PF binding. For example, SOX2 prefers to bind to its motif in one orientation more than the other, whereas OCT4 shows no preference for motif orientation[110]. Motif orientation bias can cause nucleosome occupancy levels to differ upstream vs downstream of bound TF motifs in vivo, as is the case for ELF1 and ELF2[109]. This observation could result from nucleosome repositioning upon TF binding or reflect differences in the permissiveness of motifs to TF binding relative to the underlying nucleosome position.

1.2.4 Pioneer factor binding, and activity are highly regulated and context-specific in vivo

Despite the progress in PF study over the last 20 years, currently, a debate in the field is whether all sequence-specific TFs have the potential to function as a PF within certain cellular contexts, or if pioneering ability is an intrinsic feature possessed by only a subset of TFs. A recent study investigated the pioneer factor hypothesis by comparing the pioneering activity of the canonical PF FOXA1 with that of the non-pioneering TF HNF4A by measuring the relative concentration of each factor needed to bind to accessible versus inaccessible genomic sites within naive K562 cells[113]. They found that HNF4A, previously believed to be non-pioneering, has

higher affinity for inaccessible genomic sites than FOXA1, meaning that HNF4A has stronger pioneering activity than FOXA1[113]. Many early PF studies defined PFs by their observed in vitro properties. On the other hand, the in vivo context introduces numerous cell-context-specific variables that complicate the ability to define TFs binarily as being "pioneering" or "non-pioneering" factors.

One of the facts that contrast with PFs' activity to access nucleosomal DNA is that PFs only bind to a subset of their putative binding motifs throughout the genome. A recent study in HepG2, A549 and dEN cells estimated that pioneer factor FOXA2 binds to less than 4.0% of its putative binding motifs in the genome[114]. Moreover, the binding patterns are distinct in different cell types. It has been reported that of all binding sites identified in MCF-7 and LNCaP cells (3932 sites total), only 21.7% of them (855 sites) are shared between the two cell types[77,115]. Maintenance of specific binding patterns is vital for cell functions and fate control[77,114]. Thus, understanding how binding patterns of pioneer factors are determined is an important issue. Although some relevant studies have been conducted to date[80,81,114,115], variables that affect the binding affinities and activities of pioneer factors are largely unknown.

Not only is PF binding cell-type specific, but their activity and function are also highly regulated in different contexts. Revealed by the studies of FOXA1 and PAX7, PF binding does not necessarily result in chromatin opening[114,116] in all the targets, indicating that additional factors are required for their pioneer activity in vivo. Furthermore, PFs also exhibit differential pioneer activities in different developmental stages. In Drosophila, the expression of PF Grany head (Grh) is maintained throughout developmental process, but Grh is only required for chromatin accessibility in later development[117]. These results suggest that PFs are highly regulated by features inside the cells that could be context specific.

The most prominent extrinsic features that potentially play roles include the availability of co-factors and chromatin landscape. While the ability to target and open compacted chromatin

alone has been demonstrated for some PFs *in vitro*[118], dependency on co-factors for binding and chromatin opening in vivo has also been reported recently[119-123]. On the other hand, certain epigenetic modifications have been found to enhance or inhibit the binding of pioneer factors to their binding motifs in the genome. H3K9me3 has been shown as a marker that blocks the binding of pioneer factors Sox2, Klf4 and Oct4[80]. H3K9me3 heterochromatic domains in fibroblasts are refractory to the binding by these factors. A genome-wide knocking down of H3K9me3 allows Oct4 and Sox2 to bind to these domains[80]. H3K4me2 and DNA methylation have also been reported to affect pioneer factor binding[81,115], although some other studies showed inconsistent results[124,125]. The current understanding of how co-factor availability and epigenetic landscape influence the binding specificity of PFs, as well as the contributions from this dissertation will be discussed in more detail in **Chapter 3**.

Aside from the extrinsic features discussed above, in vivo PF binding is also affected by its intrinsic properties including post-translational modifications (PTMs) and expression of different isoforms. PTMs can alter TF activities in many aspects, including DNA binding, transcriptional activation, protein degradation and subcellular localization[126]. For PFs, their unique ability to engage with chromatin could also be regulated by PTMs, exemplified by the recent studies on the methylation of lysine 270 of FOXA1 (FOXA1-K270)[127,128]. FOXA1-K270 resides at the carboxyl-end of FOXA1's DNA-binding domain and was reported to interact with core histones[91]. Methylation of FOXA1-K270 by SETD7 disrupts chromatin binding of FOXA1 globally while demethylation by LSD1 stabilizes FOXA1 binding, which subsequently alters androgen receptor binding and signaling in prostate cancer cells. Besides, other PTMs such as acetylation, phosphorylation and sumoylation also occur on FOXA1 to regulate its activity[129]. For other PFs, the roles of PTMs in their functions have been extensively studied as well[130-132].

Different isoforms of PFs may have distinct pioneering activities. For example, a recent study on the two isoforms of PAX7, which only differ by two amino acid residues in the DNA

binding paired domain, show divergent pioneer activities. While these two isoforms occupy similar genomic targets, one of them can only prime a large fraction of melanotrope-specific enhancer, but fails to fully activate them[133]. PTMs and different isoforms provide an extra regulatory layer for controlling PF binding and functions in the cells, likely in context-specific manners.

1.3 Using Synthetic DNA Library to Investigate Chromatin and Gene Regulation

1.3.1 The advantage of using synthetic DNA library

Recent progress in next-generation sequencing technologies has led to an explosion of genome-wide studies of protein binding, chromatin conformation, and gene expression. Despite enormous progress, it is still difficult to extract the genetic rules that determine factor binding, chromatin states, and gene expression level. One reason for such deficiency is that all the above are complex, multiple-variable processes. For example, TF binding is affected by motif strength, DNA shape, co-factors, and chromatin context, and transcription level is affected by promoter and enhancer strengths, nucleosome occupancy, co-factor availability, 3D genome organization, etc. In genomic measurements, effects from all these variables mix together, making it difficult to evaluate the contribution from individual variables. The sequence space required to fully explore the combinations of these variables is extremely large, far exceeding the variations provided by the native genome, especially considering evolutionary constraints.

To solve the problem above, it is important to study these processes within a controlled variable space. Ideally, there should be an experimental system where variables can be selectively perturbed one at a time while all the other variables are kept constant. This can be achieved by introducing synthetic DNA libraries into the cells. By engineering artificially designed sequences

into the same plasmid or chromatin background, these assays, in theory, allow us to measure a biological output (e.g. TF binding, chromatin accessibility, or transcription level) while changing one genetic rule at a time (e.g. number of TF motifs, motif strength, co-factor presence, +/- disease-associated mutations). Similarly, the same sequences can be integrated into different chromosome loci, making it possible to evaluate the effect from chromatin context. Depending on the experimental design, 102 – 108 synthetic sequences can be interrogated simultaneously, making it an efficient tool for genotype-phenotype mapping.

1.3.2 Using Synthetic DNA Libraries to Investigate TF Binding

As mentioned above, synthetic DNA libraries provide unique advantages in probing multivariable genetic processes. Several studies using this method have been conducted in budding yeast or mammalian cells to investigate the mechanisms that determine TF binding preference. The general experimental procedure is described in **Figure 1-9**, where DNA oligonucleotide pools containing WT or mutant TF motifs are first synthesized and cloned into a plasmid library. The plasmid library is then delivered into cells, either transiently or integrated into specific loci. TF binding to the library sequences is then measured by chromatin immunoprecipitation (ChIP)[134,135], retrotransposon insertion (or 'calling cards')[136], or DNA methylation[137,138], followed by high-throughput sequencing.

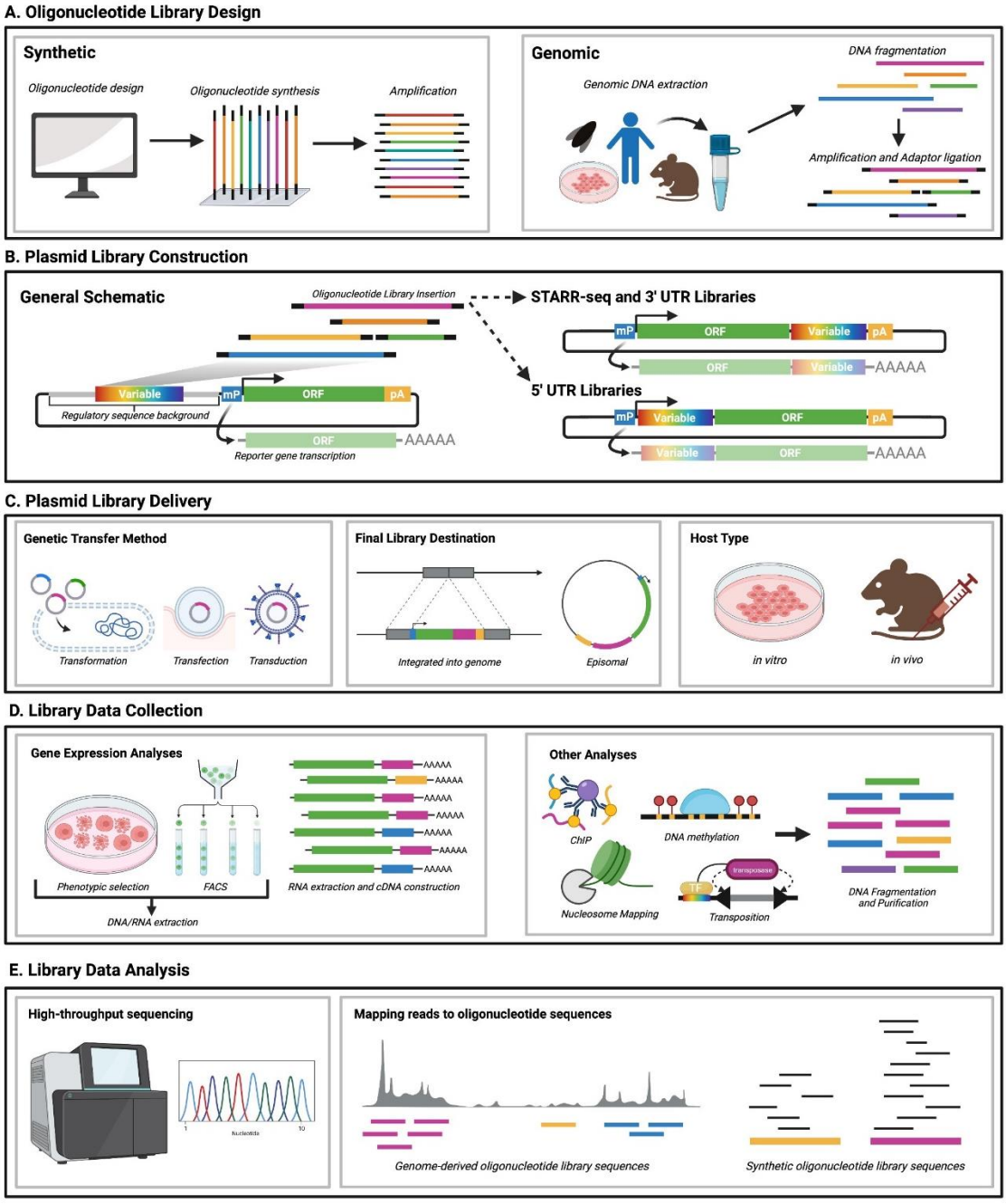


Figure 1- 9: Overview of methods employing DNA libraries to study gene regulation.

Picture from Ref[139].

These studies found that motif strength positively correlates with TF binding[134,136,137]. In addition, cooperativity and competition among TFs play a significant role in modulating TF

binding in yeast[135,136]. Incorporation of TF binding measurements into a gene expression model improves its predictive power[135]. The findings in mammalian cells are more complicated as motif occurrences tend to have low predictive power for TF binding. Two studies provide different explanations for such binding site selectivity. By swapping 25 different core motifs into 25 different flanking sequences, Grossman et al. (2017) found that in vivo binding of adipogenesis regulator PPARγ on plasmids is predominantly determined by its core motifs[134]. It was proposed that the site selectivity of PPARγ in the native genome is mainly due to differential chromatin accessibility and epigenetic modifications. Another study on Wnt effector Tcf7l2 using an integrated sequence library showed that, although local chromatin accessibility plays a role, its binding specificity is heavily affected by the 99 bp surrounding sequences[137]. In particular, the presence of Oct4 and Klf4 motifs promote Tcf7l2 binding, and the effect oscillates with the distance between Tcf7l2 and co-factor motifs with a 10.8 bp phasing, indicating the importance of interaction with co-factors at the same orientation on the DNA helix. It is possible that TFs use different strategies to achieve binding specificity. Studies on more TFs need to be done to see if there is a dominant strategy.

A different experimental design by Vanzan et al. (2021) used an indirect method to infer pioneer transcription factor (PF) binding to DNA by measuring differential methylation status and screened reported PFs for their ability to induce methylation changes[140]. In this study, DNA libraries containing binding motifs of different mammalian TFs are either methylated in vitro or left unmethylated and integrated into the mammalian genome. Changes in methylation levels after integration are then measured to infer the binding and effect of the corresponding PFs. The results revealed two groups of PFs: protective PFs (PPFs) which protect DNA from methylation and super PFs (SPFs) which induce DNA demethylation at methylated binding sites.

Chapter 2

Systematic Dissection of Sequence Features Affecting the Binding Specificity of a Pioneer Factor Reveals Binding Synergy Between FOXA1 and AP-1

Work presented in this chapter is a part of the manuscript preprint "Xu, C.*, Kleinschmidt, H.*, Yang, J., Leith, E., Johnson, J., Tan, S., Mahony, S., & Bai, L. (2023). Systematic Dissection of Sequence Features Affecting the Binding Specificity of a Pioneer Factor Reveals Binding Synergy Between FOXA1 and AP-1. (* Contributed equally). bioRxiv, https://doi.org/10.1101/2023.11.08.566246.".

2.1 Abstract

Despite the unique ability of pioneer transcription factors (PFs) to target nucleosomal sites in closed chromatin, they only bind a small fraction of their genomic motifs. The underlying mechanism of this selectivity is not well understood. Here, we design a high-throughput assay called ChIP-ISO to systematically dissect sequence features affecting the binding specificity of a classic PF, FOXA1. Combining ChIP-ISO with in vitro and neural network analyses, we find that 1) FOXA1 binding is strongly affected by co-binding TFs AP-1 and CEBPB, 2) FOXA1 and AP-1 show binding cooperativity in vitro, 3) FOXA1's binding is determined more by local sequences than chromatin context, including eu-/heterochromatin, and 4) AP-1 is partially responsible for differential binding of FOXA1 in different cell types. Our study presents a framework for elucidating genetic rules underlying PF binding specificity and reveals a mechanism for context-specific regulation of its binding.

2.2 Introduction

Sequence-specific transcription factors (TFs) are major regulators of gene expression. Characterization of the location and strength of TF binding in the genome is therefore a critical step in understanding gene regulation. TF binding sites are typically identified using the weight matrices of their binding motifs. In higher eukaryotes, however, this method has weak predictive power for actual TF binding events. Many TFs bind <1% of their motifs across the genome, and their binding patterns can change in a cell-type-specific manner[141-143]. Multiple features beyond the core sequence motif have been proposed to contribute to this phenomenon, including DNA shape[144,145], cooperative binding with other TFs[66,114,146], DNA methylation[68,147], nucleosome occupancy and chromatin accessibility[69,148], histone modifications[70,80], 3D genome contacts[71], and variations in local TF concentrations[72,149]. Among these potential factors, nucleosomes have a major inhibitory effect on the binding of many TFs, and chromatin accessibility is therefore considered to be the key determinant of TF binding[148,150-152].

A subset of TFs known as "pioneer factors" (PFs) can stably associate with nucleosomal templates by recognizing partial sequence motifs and/or interacting with histones[70,88,92,103,153]. Inside cells, PFs can overcome the nucleosomal barrier by targeting nucleosome-embedded motifs and generating accessible chromatin, which enables the binding of other TFs and triggers transcriptional activation[76,93,154,155]. Given their ability to open chromatin in vivo and bind nucleosomal DNA in vitro, PFs should be able to access most, if not all, consensus motifs in the genome. This is indeed the case for PFs in budding yeast[156] (**Figure 2-1**). Surprisingly, like canonical TFs, PFs in higher eukaryotes also show highly selective and cell-type specific binding. For example, FOXA1 is a classic PF capable of binding and opening highly compacted chromatin[118,157-159], but it only occupies 3.7% of its potential motifs in MCF-7 cells, and less than half of these binding events are shared with LNCaP cells[115]. Our analysis found that only 10-20%

consensus motifs are bound by FOXA1 in MCF-7 and A549 cells (**Figure 2-2**). The molecular mechanism underlying such binding selectivity is not well understood.

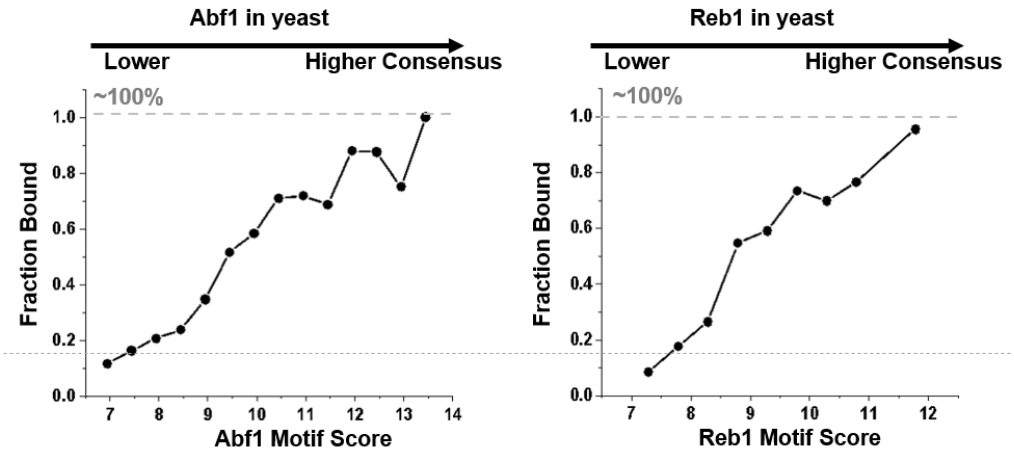


Figure 2- 1 . PFs in yeast bind to a large fraction of their consensus.

Bound fraction as a function of motif score for yeast PFs, Abf1 (left) and Reb1 (right). Dotted line represents the fraction of perfect consensus motifs that are occupied by the corresponding PFs.

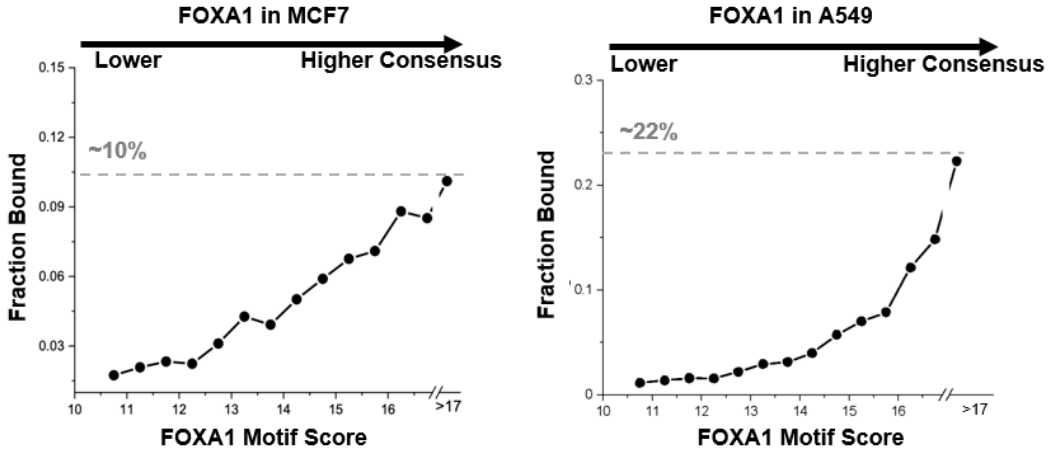


Figure 2- 2 . PFs in mammalian cells bind to a small fraction of their consensus.

Same as Figure 2-1, but for human pioneer factor FOXA1 in MCF-7 cells (left) and A549 cells (right).

TF binding is usually studied in the context of the native genome, where each binding event can be affected by multiple variables, and individual effects are therefore hard to dissect.

Here, we sought to overcome this limitation by developing a new method named “Chromatin Immunoprecipitation with Integrated Synthetic Oligonucleotides (ChIP-ISO)” and applied this method to study FOXA1 binding in human A549 lung cancer cells. In this method, we engineered specific genetic features into synthetic sequences, integrated them into a fixed genomic locus, and measured FOXA1 binding in this highly controlled genetic and epigenetic context. In combination with in vitro and neural network analyses, our work reveals key determinants of PF binding, which has implications on PF function through development and differentiation

2.3 Results

2.3.1 ChIP-ISO assay allows highly parallel measurements of FOXA1 binding to thousands of integrated synthetic sequences.

FOXA1 is expressed in A549 cells, where it plays important physiological roles[160,161]. To study FOXA1 binding specificity in this cell line, we performed the ChIP-ISO procedure as shown in **Figure 2-3**. Briefly, a synthetic oligo library was inserted into a plasmid backbone to generate a plasmid library, which was then integrated into the AAVS1 locus in the human genome through CRISPR-Cas9, and FOXA1 binding to these sequences was measured by ChIP followed by amplicon sequencing.

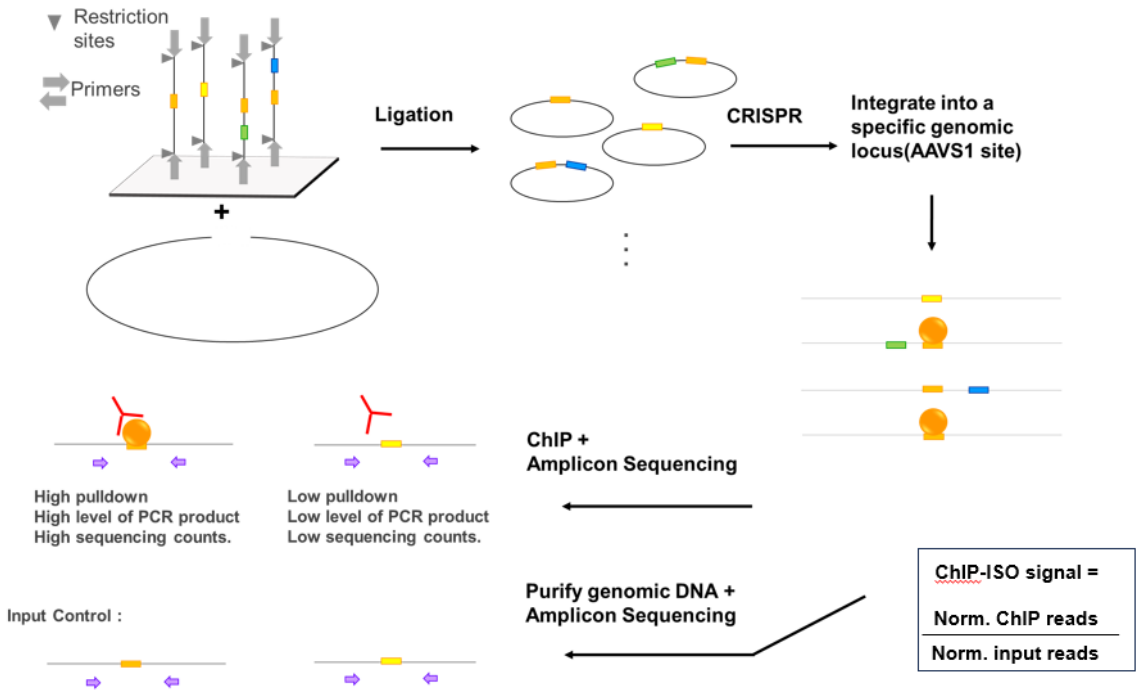


Figure 2- 3: Workflows for ChIP-ISO.

See main text for detailed steps of ChIP-ISO.

The synthetic oligonucleotide library used here contains 3,203 different sequences that are 229-bp in length, including a 193-bp variable region (**Figure 2-4**). These sequences include fragments from the native genome with FOXA1 motifs and their variants, which can be divided into three categories, each aiming to test different variables that potentially affect FOXA1 binding (Table 2-1). The library was synthesized, PCR amplified, digested, and ligated into a plasmid backbone containing the flanking sequences of the CCND1 enhancer (CCND1e)[162] (CCND1eΔ), in which all the endogenous FOXA1 motifs are deleted (**Figure 2-5**)

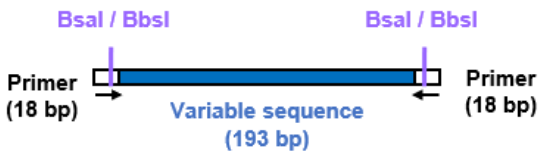


Figure 2- 4 Diagram of ChIP-ISO library oligonucleotide design.

The variable region of each sequence spans 193 bp, as indicated in blue. The BsaI / BbsI is the restriction enzyme recognition sites engineered in the 18 bp flanking primer sequences.

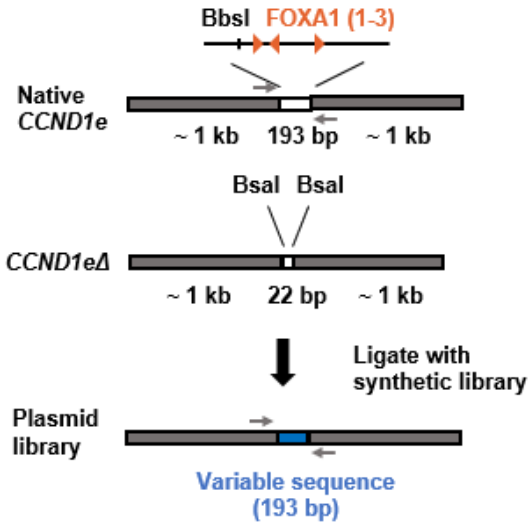


Figure 2- 5 Plasmid construction strategy.

We start off with the native CCND1 enhancer sequence, cloned into the pAAVS1-Nst-MCS plasmid background. We then delete a 193 bp region containing all three FOXA1 binding sites, as well as a BbsI cutting site, and replace it with two BsaI cutting sites. Finally, we ligate the 193 bp ChIP-ISO oligonucleotides (blue) into this plasmid using the BsaI sites. Grey arrows represent primers used for ChIP-ISO amplicon sequencing.

Table 2- 1: Library design.

Table summarizing the three subsets of our ChIP-ISO oligonucleotide library.

Set	Description	No. of seqs
1	*CCND1* enhancer and its variants with scanning mutations, motif mutation, etc.	560
2	Co-bound sites from the genome and their variants with FOXA1 or co-factor motifs mutated	791
3	Sites from different genomic background and their variants with FOXA1 motifs mutated	1852

The resulting plasmid library was then transfected into A549 cells and integrated into the AAVS1 safe harbor locus[163] through CRISPR-Cas9. We used an engineered Cas9 nuclease[164] and a promoter-less selection maker[165] to reduce potential off-target integration (**Figure 2-6**).

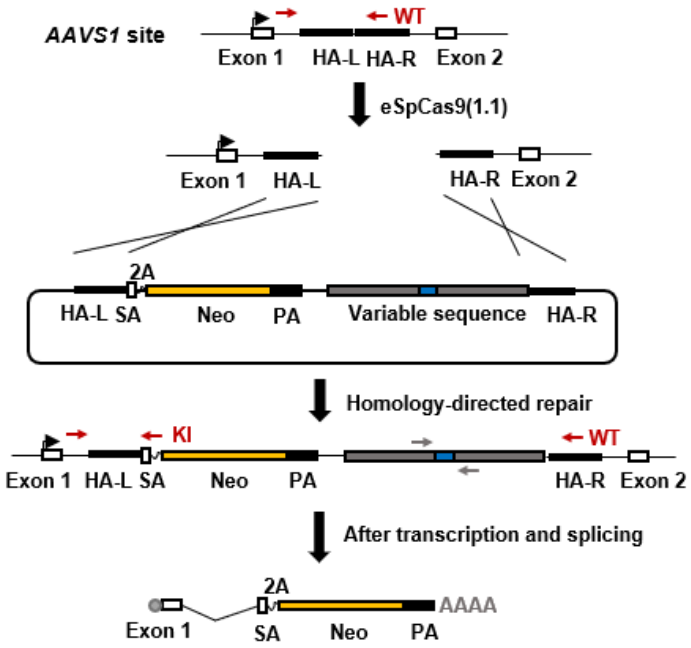


Figure 2- 6 Strategy for integration of the plasmid library into the AAVS1 site using CRISPR/Cas9.

Double-strand break is generated at the AAVS1 site by pSpCas9 (1.1). The region containing library and background sequences, as well as the selection marker (Neo), is integrated into the AAVS1 site (inside the first intron of PPP1R12C gene) by homology-directed repair. If integrated successfully, the promoterless Neo gene will be transcribed together with the endogenous PPP1R12C gene, and RNA splicing will happen between exon 1 and the splicing acceptor (SA) site. A 2A peptide is engineered before the Neo gene to make sure the translated Neo protein is folded independently. HA-L, left homologous arm; HA-R, right homologous arm. WT, a primer pair amplifying unintegrated AAVS1 site; KI, a primer pair amplifying AAVS1 site with successful integration.

We obtained a cell library with ~92k colonies (29 colonies per sequence, on average), and random genotyping showed that >80% of the colonies have the synthetic sequences integrated into the correct locus (**Figure 2-7&2-8**).

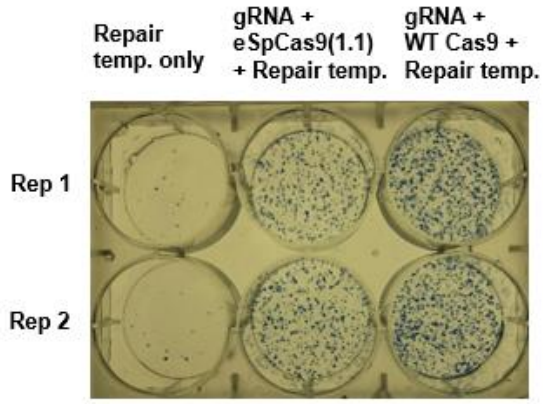


Figure 2- 7: A549 colonies after integration with no Cas9 (left, control), eSpCas9, and wt Cas9. The colonies are selected with G418 for 12 days, and stained with methylene blue.

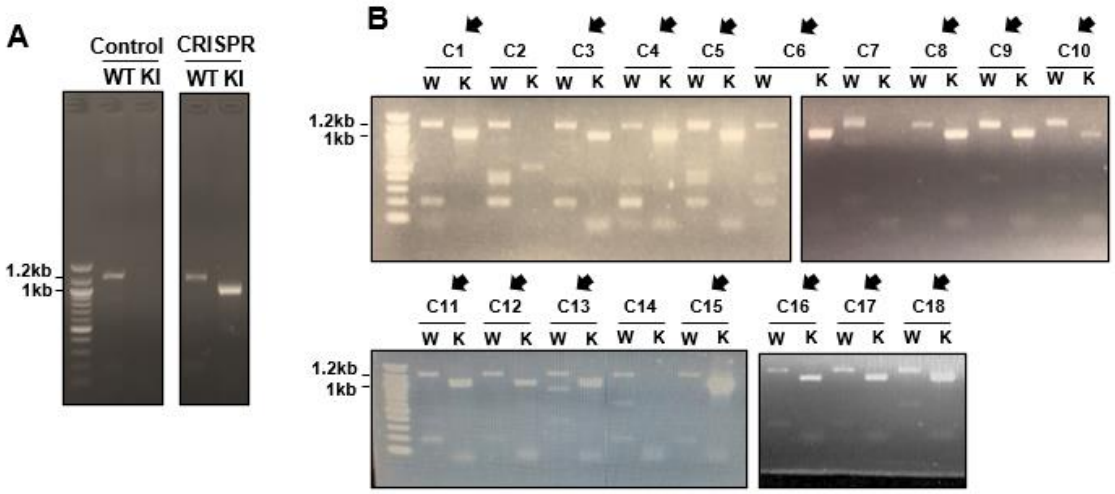


Figure 2- 8: 15 out of the 18 colonies show the right PCR band.

A) PCR test on mixed colonies after integration. The "KI" primer pair amplifies across the integration junction, and therefore the band is only visible when the plasmid sequence is integrated into the right locus. B) Same as in panel A except that the PCR is carried out in 18 single colonies. 15 out of the 18 colonies show the right PCR band. Data collected by Jenna Johnson.

FOXA1 ChIP was conducted on a mixed culture of the cell library. Genomic DNA was extracted from the same library for the input control. The region containing the synthetic sequences was PCR amplified from both samples and subjected to next-generation amplicon sequencing (**Figure 2-9**).

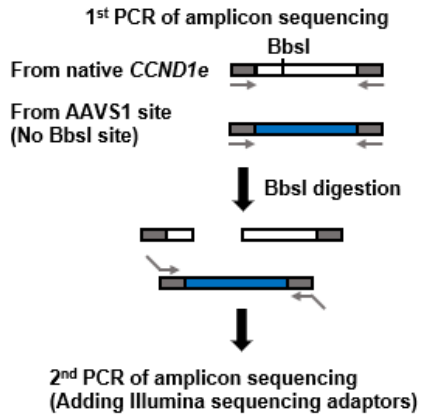


Figure 2- 9: Strategy of amplicon sequencing.

After first round of PCR using the primer pair in Figure 2-5, we perform BbsI digestion, which cuts the native CCND1 enhancers, but not the synthetic ones. The second round PCR (with sequencing adaptors) can therefore selectively amplify the ISO library for sequencing.

Since the synthetic sequences associated with FOXA1 are enriched by ChIP, for each library sequence, the ratio of the read count in the ChIP sample divided by that in the input sample was used as a measure of FOXA1 binding strength (referred to as the "ChIP-ISO signal" below).

We performed a few tests to evaluate the ChIP-ISO method. Two biological replicates agree well with an overall correlation coefficient of 0.76 (**Figure 2-10A**). The ChIP-ISO signals follow a bimodal distribution, with 56.3% of the sequences in the lower peak, representing no or low-level FOXA1 binding (**Figure 2-10B**). As expected, sequences with mutated FOXA1 motifs show lower binding (**Figure 2-10A**). Furthermore, we constructed three cell lines each containing a single library sequence integrated into the AAVS1 site, and measured FOXA1 binding by ChIP followed by quantitative PCR (ChIP-qPCR). These low-throughput measurements agree well with the high-throughput results (**Figure 2-10C**). We therefore conclude that the ChIP-ISO method can accurately and efficiently measure FOXA1 binding to integrated synthetic sequences.

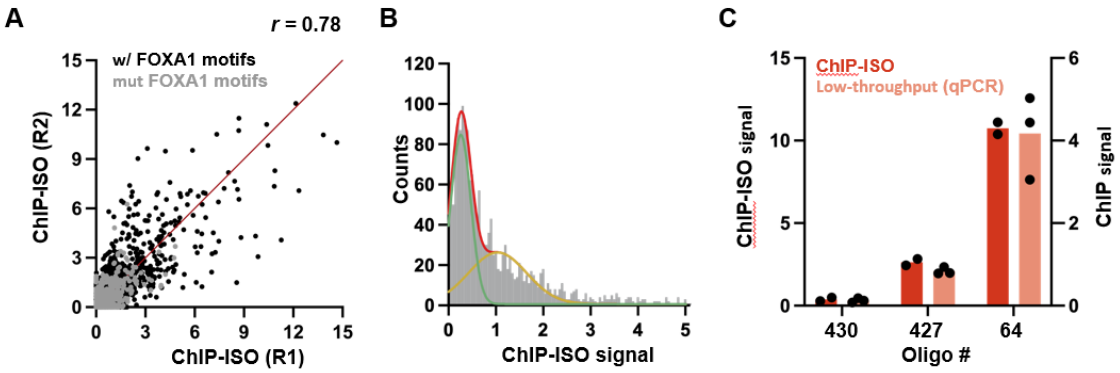


Figure 2- 10: ChIP-ISO assay allows highly parallel measurements of FOXA1 binding to thousands of integrated synthetic sequences.

A) Reproducibility of FOXA1 ChIP-ISO signal across two biological replicates. Black / gray dots represent sequences containing WT / mutated FOXA1 motifs. r: Pearson correlation. B) Histogram of the FOXA1 ChIP-ISO signals with the entire ISO library, fit by two Gaussian peaks (green and yellow: low and high peaks, respectively; red: superposition of the two). C) Comparison of FOXA1 ChIP-ISO signals with low-throughput ChIP-qPCR signals from three biological replicates over three individual ISO library sequences.

2.3.2 Co-binding of AP-1 strongly enhances FOXA1 binding to the CCND1 enhancer.

The endogenous CCND1e in A549 cells is bound by FOXA1 and accompanied by high chromatin accessibility and H3K27ac signals[162] (**Figure 2-11**). It contains three FOXA1 motifs, as well as conserved binding sites of eight other TFs (**Figure 2-12&2-13**).

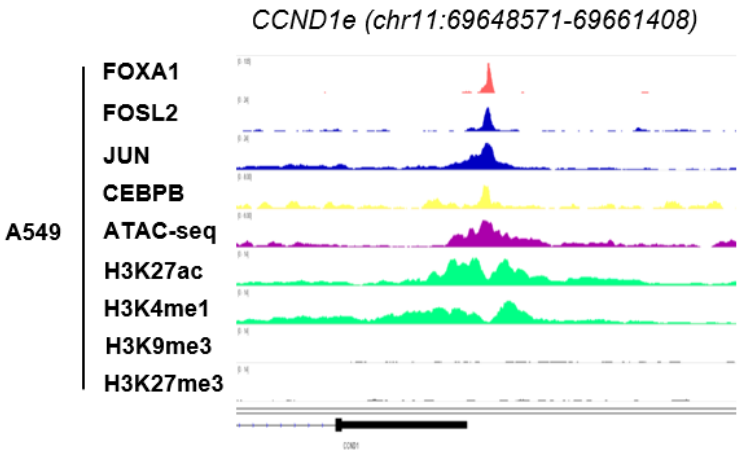


Figure 2- 11: Genomic tracks of TF and histone modification ChIP-seq and ATAC-seq signals at the CCND1 enhancer in WT A549 cells.

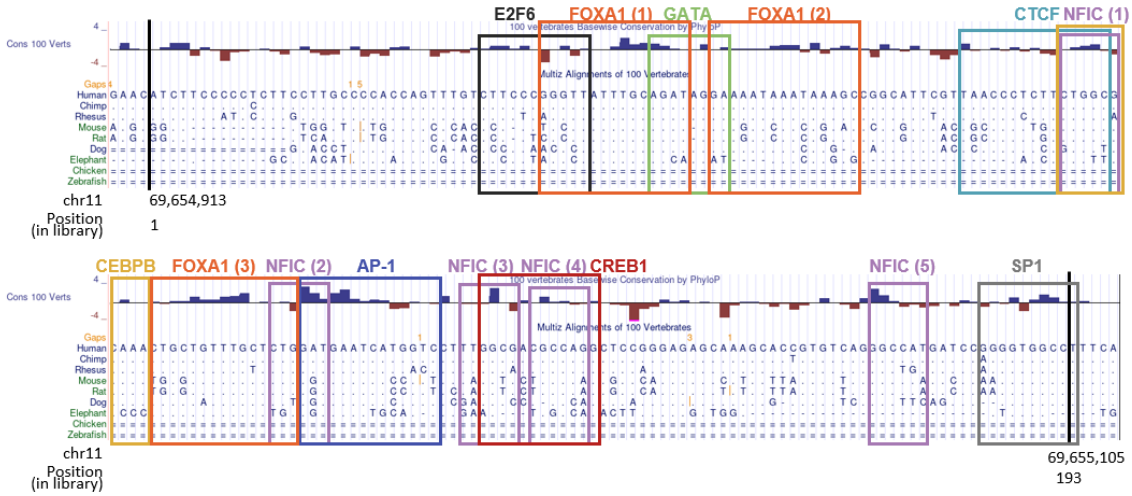


Figure 2- 12: Conservation of DNA sequence within the ~200 bp CCND1 enhancer.

Blue: conserved nucleotide, brown: non-conserved. TF motifs are outlined and color coded. Vertical black lines marks the exact 193 bp sequences use in our ISO library.

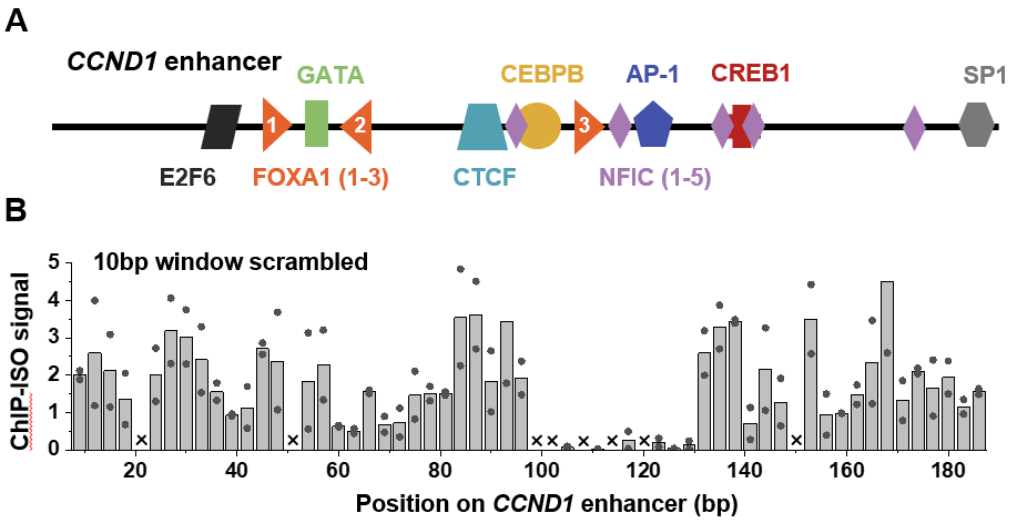


Figure 2- 13: ChIP-ISO signals over CCND1e variants containing scrambled sequences in a 10bp moving window (step size: 3bp).

A) Map of the 193bp portion of the CCND1e explored in this study (chr11:69,654,913-69,655,105). Three FOXA1 motifs (orientations depicted by arrow directions) and motifs of potential co-binding TFs are labeled. B) Bar: averaged ChIP-ISO signal; Dot: data from individual replica; X: missing data. Gray box highlights an area where the scrambles lead to particularly low FOXA1 ChIP-ISO signals, indicating that these sequences are critical for FOXA1 binding.

The first set of the library includes CCND1e mutants. **Figure 2-13** shows the ChIP-ISO measurement on sequences with scanning mutations, where a 10bp window is sequentially scrambled with a 3bp step size. The most prominent drop in FOXA1 occupancy is observed when a region near the third FOXA1 motif is scrambled, indicating that this region contains key elements that recruit FOXA1.

To more accurately pinpoint sequence features affecting FOXA1 binding, we first replaced each FOXA1 motif with mutated, reversed, or consensus versions. As expected, mutating/strengthening FOXA1 motifs significantly reduce/increases FOXA1 binding, respectively, while reversing their orientation has a minor effect (**Figure 2-14**). Consistent with **Figure 2-13**, the third FOXA1 motif is more influential than the other two (**Figure 2-14&2-15**).

FOXA1 binding also increases with the number of FOXA1 motifs in a non-linear fashion (**Figure 2-15**), indicating that there is a synergistic effect among multiple adjacent binding sites.

FOXA1 Motif			FC (log_2)	*P*-value ($-log_{10}$)
1st		mut	-0.33	1.00
		rev	0.12	0.27
		con	0.96	2.26
2nd		mut	-0.54	1.74
		rev	-0.48	1.25
		con	0.98	3.25
3rd		mut	-1.81	3.49
		rev	-0.74	2.30
		con	1.02	4.59

Figure 2- 14: The effect on FOXA1 binding by manipulating individual FOXA1 motifs.

Each FOXA1 motif is either mutated (mut), orientation-reversed (rev), or converted into the strongest consensus (con). The table lists the fold change and statistical significance of FOXA1 ChIP-ISO signal caused by these sequence variations. Two-tailed paired t-test.

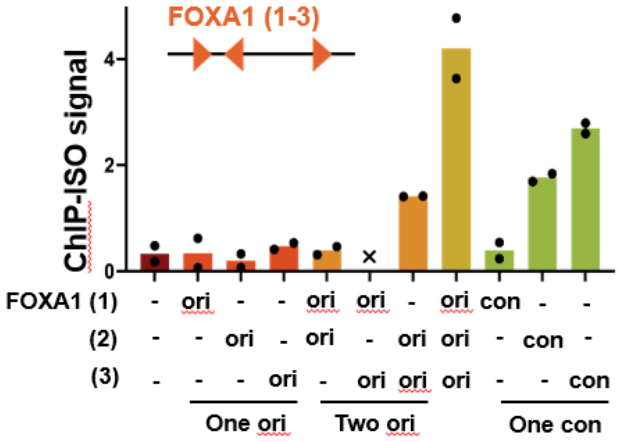


Figure 2- 15: Example FOXA1 ChIP-ISO signals over CCND1e variants containing 0, 1, 2, or 3 original (ori) FOXA1 motifs or 1 consensus (con) FOXA1 motif.

X: missing data

We next examined the effect of other TFs that potentially co-bind with FOXA1. As there are eight co-factor motifs, the ChIP-ISO library includes all 256 combinations where each motif can be wild-type (WT) or mutated. We found that mutating AP-1 and CEBPB motifs leads to a significant decrease in FOXA1 binding (**Figure 2-16**). AP-1 has a particularly strong effect, as mutating its motif dramatically decreases FOXA1 binding to a level close to the background for almost all CCND1e variants (**Figure 2-17**)

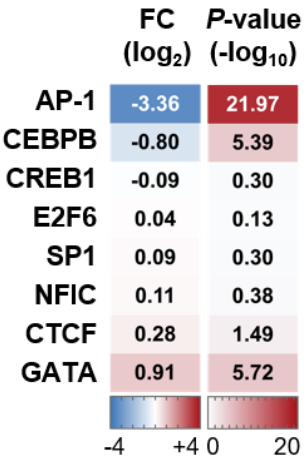


Figure 2- 16: The effect on FOXA1 binding by mutating co-factor motifs.

The table lists the fold change and statistical significance of FOXA1 ChIP-ISO signal caused by these sequence variations. Two-tailed paired t-test.

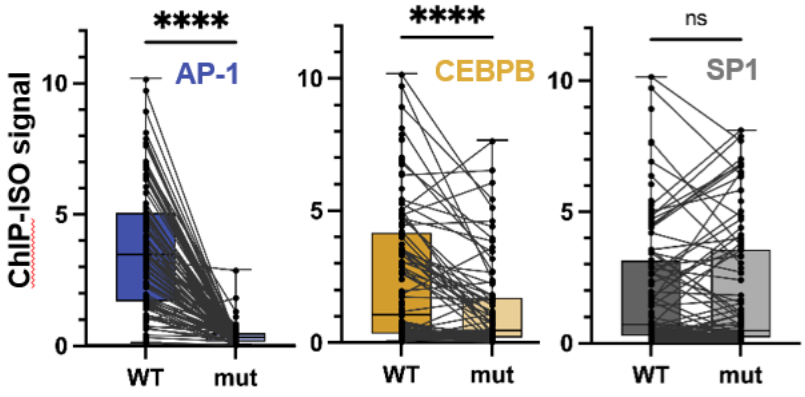


Figure 2- 17: Box-and-whisker plots showing FOXA1 ChIP-ISO signals over otherwise identical CCND1e sequences containing WT or mutated AP-1, CEBPB, or SP1 motifs.

Paired sequences are connected by a line. ****: $p < 0.0001$, ***: $p < 0.001$, **: $p < 0.01$, and ns: non-significant based on two-tailed paired t-test (same below, unless specified).

The presence of AP-1 highly correlates with FOXA1 binding, even more than the total FOXA1 motif score (**Figure 2-18**). Low-throughput FOSL2 (a subunit of AP-1) and FOXA1 ChIP confirmed the abolished binding of both factors when the AP-1 motif is mutated (**Figure 2-19**). This data indicates that AP-1 is a crucial co-factor that potentiates FOXA1 binding to the CCND1e. Interestingly, both AP-1 and CEBPB motifs are immediately adjacent to the third FOXA1 motif, which has the largest impact on FOXA1 binding (**Figure 2-13, 2-14&2-15**).

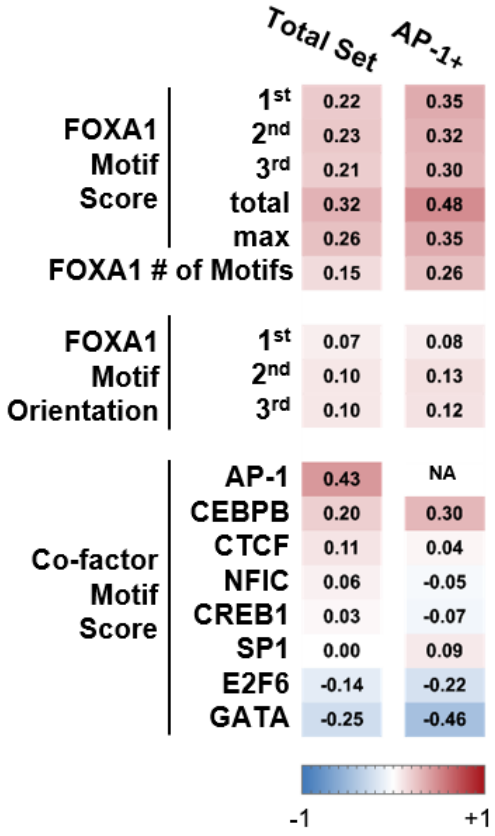


Figure 2- 18: Pearson correlation coefficient between FOXA1 ChIP-ISO signals and CCND1 sequence variables.

Calculated using the total set or the subset containing an AP-1 motif. These numbers reflect the level of impact of each variable on FOXA1 binding.

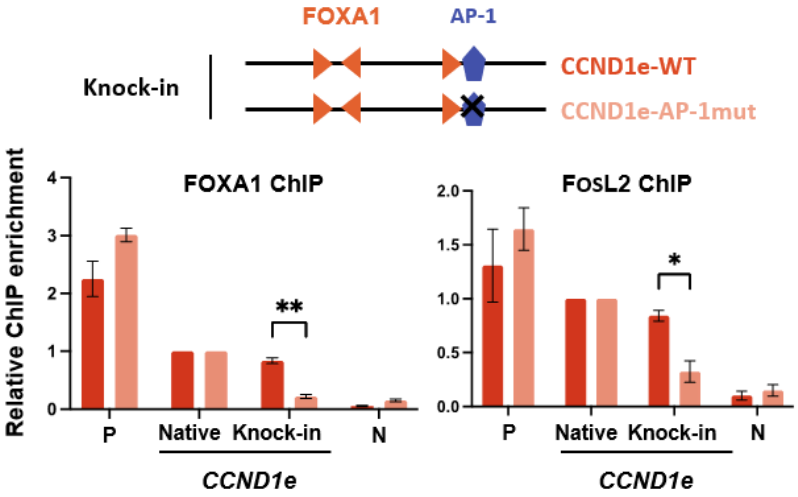


Figure 2- 19: Low-throughput FOXA1 and FOSL2 ChIP-qPCR for three biological replicates showing the impact of the AP-1 motif on their binding.

P: positive control, Native: native CCND1 enhancer, Knock-in = AAVS1-integrated CCND1 enhancer (wt or AP-1 mut), N: negative control. *: $p < 0.05$, **: $p < 0.01$.

To test the significance of this observation, we moved the AP-1 motif away from the third motif. We found that FOXA1 binding declines markedly with increasing distance (**Figure 2-20**), indicating that motif proximity is important for AP-1 facilitated FOXA1 binding.

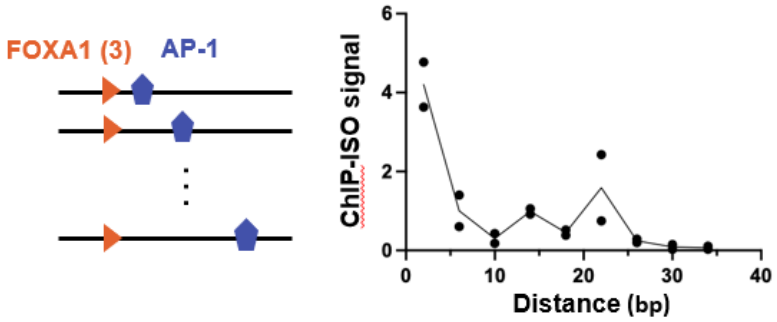


Figure 2- 20: FOXA1 ChIP-ISO signal as a function of the linear distance between the 3rd FOXA1 motif and the AP-1 motif.

Correlation analysis shows that FOXA1 binding is mostly affected by AP-1, its own motif, and CEBPB (**Figure 2-18**). To understand the interplay between these factors, we plotted

the FOXA1 binding strength as a function of the total FOXA1 motif score in the presence or absence of AP-1 and CEBPB (**Figure 2-21**). Without AP-1 and CEBPB, FOXA1 can still bind strong motifs, but the presence of these two co-factors allows FOXA1 to target sub-optimal motifs, at least in the CCND1e context (**Figure 2-21**).

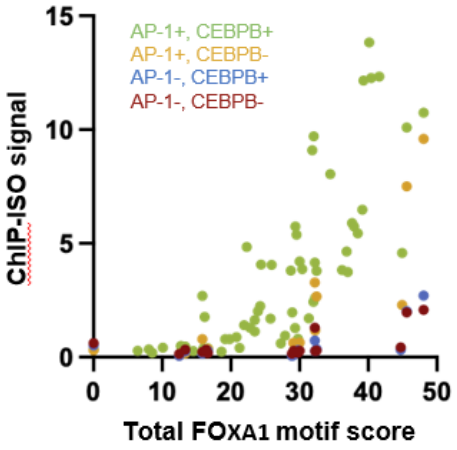


Figure 2- 21: Relation between the FOXA1 ChIP-ISO signal and the total FOXA1 motif score over CCND1 variants ± AP-1 / CEBPB motifs.

We also investigated the reciprocal relationship of FOXA1 on AP-1 binding to determine if AP-1 binds upstream of FOXA1 or if they bind cooperatively to enhance each other's binding. We addressed this question by generating a new cell line containing the CCND1e with all three FOXA1 motifs mutated and measured AP-1 binding in the absence of FOXA1 with ChIP-qPCR. Notably, both FOXA1 and AP-1 binding are drastically reduced on this mutated CCND1e (**Figure 2-22**), supporting the scenario that the binding of these two TFs is mutually dependent and cooperative.

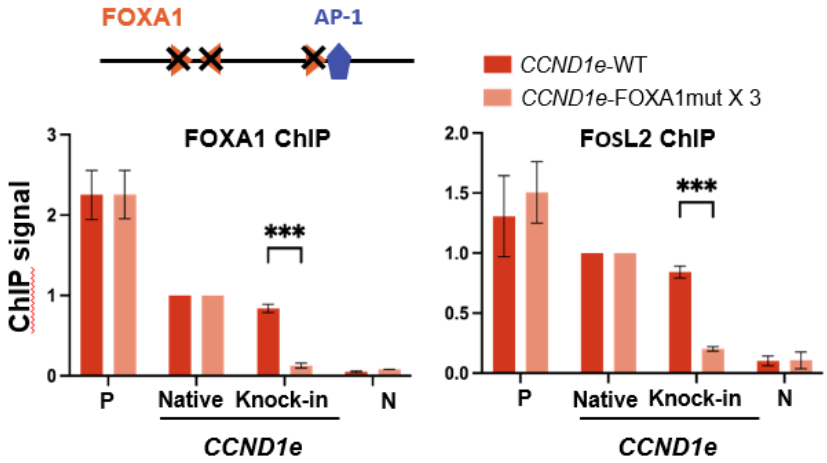


Figure 2- 22: Effect of FOXA1 on AP-1 binding.

Bar plots show FOXA1 and FOSL2 ChIP-qPCR for three biological replicates over integrated WT CCND1e or a variant with all three FOXA1 motifs mutated (Knock-in). ChIP-qPCR over a positive / negative control locus (P and N) and the native CCND1e are also shown. Error bars represent standard error (same below, unless specified).

2.3.3 AP-1 and CEBPB co-bind with FOXA1 and assist its binding genome-wide.

The case study of the CCND1e demonstrates the importance of co-factors in FOXA1 binding. We next asked if this phenomenon applies to other genomic loci and/or with other co-factors. We first evaluated the co-binding of FOXA1 with other TFs in A549 cells based on the overlap between their ChIP-seq peaks and the occurrence of their motifs in FOXA1 peaks. A large fraction (25% to 45%) of FOXA1 ChIP-seq peaks overlap with the peaks of AP-1 subunits JUNB, JUND, and FOSL2, and vice versa (**Figure 2-23**). Moreover, the most enriched motifs within FOXA1 peaks, aside from the FOXA1 motif itself, are those of the AP-1 subunits (P-value < 10-1000) (**Figure 2-23**). This data supports wide-spread co-binding of FOXA1 and AP-1. Many other TFs, including CEBPB, also display significant co-binding with FOXA1 (**Figure 2-23**).

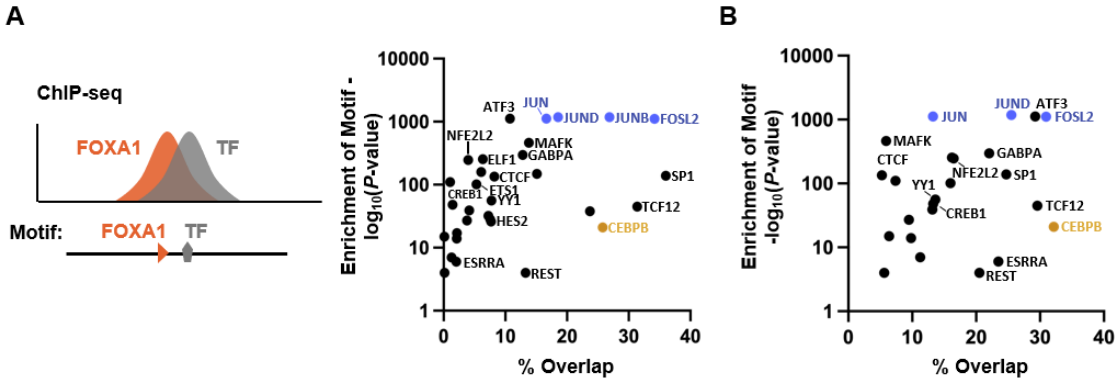


Figure 2- 23: Bioinformatic analysis of FOXA1/TF co-binding.

A) Schematics of co-binding events. For each TF, the dot plot shows the percentage of the overlapped FOXA1 ChIP-seq peaks (x axis) and the enrichment of its motif within FOXA1 peaks (y axis). TFs chosen for further analysis are labeled. AP-1 subunits are indicated in blue. B) The same as A right panel, except the x axis represents the percentage of a TF ChIP-seq peaks overlapped with FOXA1 peaks (A is the percentage of the overlapped FOXA1 peaks). Data collected by Holly Godin.

Given that co-factors may permit FOXA1 binding at suboptimal motifs (**Figure 2-21**), we analyzed FOXA1-TF co-binding at FOXA1 sites with different motif strengths. We separated FOXA1 binding events near strong consensus motifs (scores > 16) or very weak ones (scores < 12) (**Figure 2-24**, left column). The average FOXA1 binding strength is comparable over these two sets of regions. Strikingly, co-binding predominantly occurs in the sites with low motif scores (**Figure 2-24**). These sites also show active histone marks and high chromatin accessibility (**Figure 2-24**). This data suggests that TF "hubs" tend to form over weaker motifs, and these co-binding events are more likely to be functional in gene regulation. This may represent a common strategy to ensure gene expression plasticity (see **Discussion**).

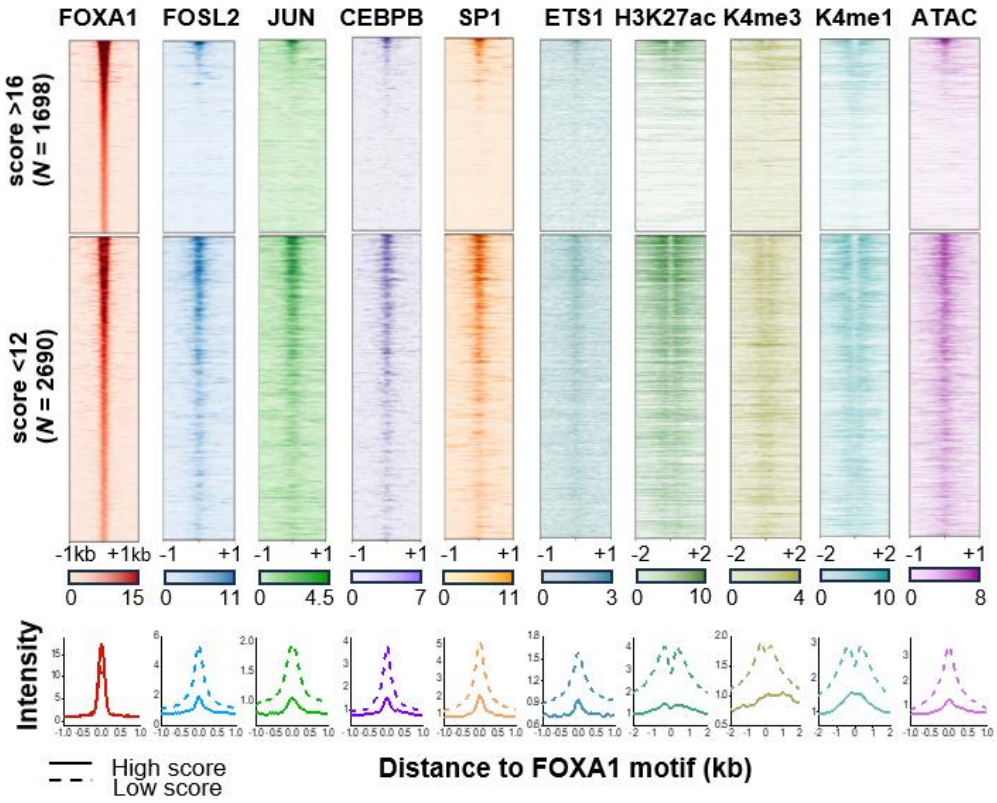


Figure 2- 24: Heatmaps of ChIP-seq / ATAC-seq and the corresponding intensity profiles in WT A549 cells over FOXA1 binding regions separated based on FOXA1 motif scores.

Sequences in the top section contain strong consensus motifs (score >16), and the ones at the bottom contain weak motifs (score < 12).

Co-binding between FOXA1 and TFs does not necessarily imply cooperativity among these factors. To test if the TFs identified in **Figure 2-23** indeed affect FOXA1 binding, we carried out additional mutational analyses for 15 TFs that show co-binding with FOXA1. For each TF, we selected 10-20 native genomic loci where it co-occupied by FOXA1 with proximal motifs (<30bp). The 193bp sequences from these loci, together with variants containing mutated TF motifs, were included in the ChIP-ISO library to evaluate the impact of these motifs on FOXA1 binding. AP-1 motif mutation again has the largest impact on FOXA1 binding, followed by CEBPB (**Figure 2-25&2-26**), indicating that these two factors promote FOXA1 binding at

many genomic loci. These results also suggest that most co-localized TFs do not bind cooperatively with FOXA1.

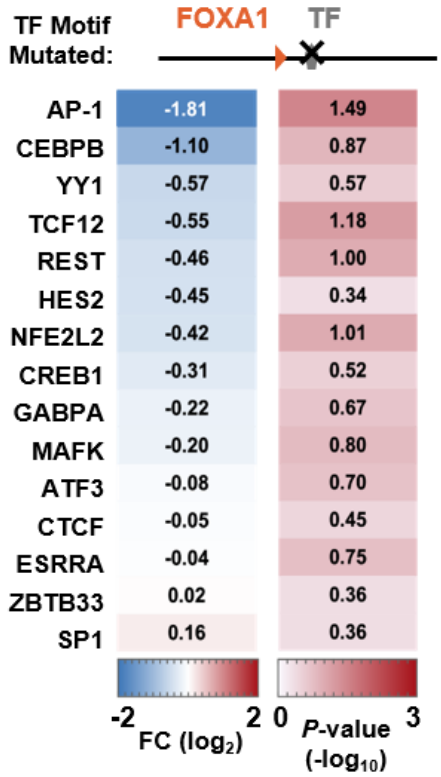


Figure 2- 25: The effect on FOXA1 binding by mutating co-factor motifs.

The ISO set is derived from genomic sequences that show overlapped FOXA1 and TF ChIP-seq peaks in A549 and contain both motifs in proximity (<30bp). The table lists the fold change and statistical significance of FOXA1 ChIP-ISO signal upon mutations of co-factor motifs in these sequences. Two-tailed paired t-test.

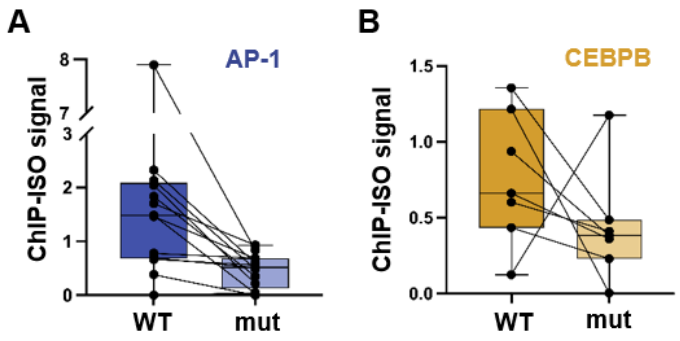


Figure 2- 26: Box-and-whisker plots showing the changes of FOXA1 ChIP-ISO signal upon mutations.

Mutations of A) AP-1 or B) CEBPB motif.

To further assess whether co-factor motifs are predictive of genome-wide FOXA1 binding in A549 cells, we trained a convolutional neural network (CNN) to recognize FOXA1 ChIP-seq peaks using DNA sequence features. The CNN achieves high overall performance, with an area under precision-recall curve of 0.66 (calculated on held-out test sites). The CNN predictions of FOXA1 binding activities are generally consistent with ChIP-ISO measurements (**Figure 2-27**). Using the DeepLift-SHAP feature attribution approach[166,167], we characterized which DNA base positions contribute towards the CNN's FOXA1 binding predictions at specific loci. Over CCND1e, for example, DeepLift-SHAP strongly highlights the second and third FOXA1 motifs and the AP-1 motif as positive contribution to FOXA1 binding (**Figure 2-28A**). The feature attribution scores at the FOXA1, AP-1, and CEBPB motifs are weakened when mutated, and those from FOXA1 are strengthened when replaced with the consensus, consistent with ChIP-ISO measurements. In addition, we ran the TF-MoDISco tool[16840] to compile feature attribution scores from across all ChIP-seq peaks into commonly occurring motif patterns. Alongside cognate FOXA1 binding motif variants, TF-MoDISco identifies the AP-1 and CEBPB motifs as the most prominent co-factor motifs that the CNN uses to predict FOXA1 binding (**Figure 2-28B**). A GC-rich sequence similar to the SP1 motif is identified as the most negative

feature (**Figure 2-28B**). In summary, our CNN analysis of FOXA1 ChIP-seq data is consistent with our findings that AP-1 and CEBPB assist FOXA1 binding genome-wide in A549 cells.

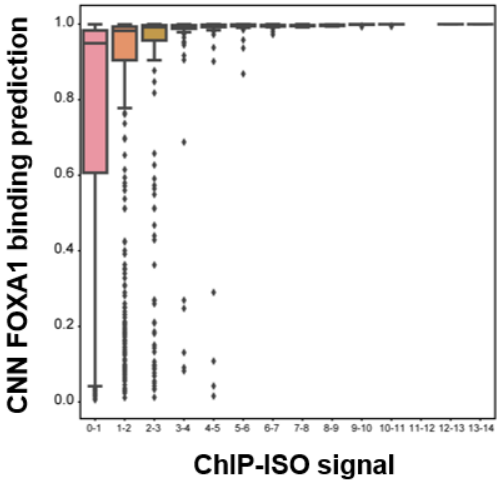


Figure 2- 27: Histogram showing distributions of scores given by a sequence-trained CNN to sequences that were tested by ChIP-ISO.

The sequences are grouped according to ChIP-ISO signal. The CNN is trained to predict FOXA1 ChIP-seq data in A549 cells. Data collected by Jianyu Yang.

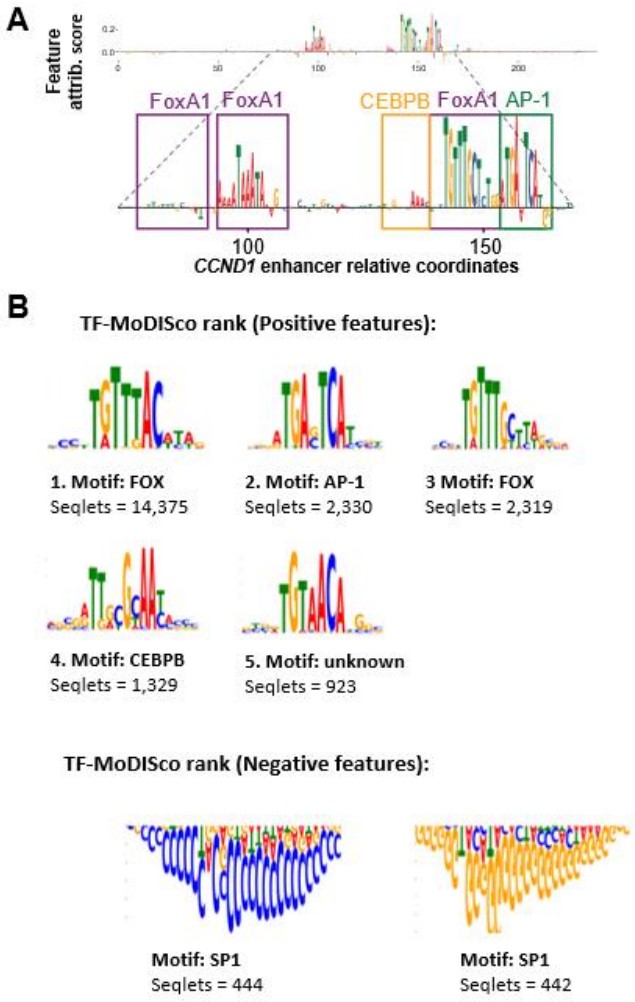


Figure 2- 28: DeepLIFT-Shap feature attribution scores highlighting features used by a sequence-trained CNN to predict FOXA1 binding in A549 cells at the CCND1e.

A) Feature attribution scores at the CCND1e. B) Top 5 motifs detected by TF-MoDISco in the genome-wide DeepLIFT-Shap positive feature attribution scores at sites predicted by the CNN to be bound by FOXA1. The TF family of matching motifs is annotated for each motif, as is the number of seqlets used by TF-MoDISco to construct each motif. Data collected by Jianyu Yang.

2.3.4 AP-1 inhibition leads to motif-directed redistribution of FOXA1 binding in the genome.

To further test the role of AP-1 in promoting FOXA1 binding, we measured the effect of knocking down AP-1 on genome-wide FOXA1 binding. Since the AP-1 family has multiple homologs that may have redundant functions, we took advantage of a dominant-negative protein A-FOS to inhibit global AP-1 binding. A-FOS dimerizes with JUN family proteins to form a heterodimer that cannot bind DNA[169,170]. We constructed an A549 cell line with doxycycline (Dox)-inducible A-FOS expression (**Figure 2-29&2-30**).

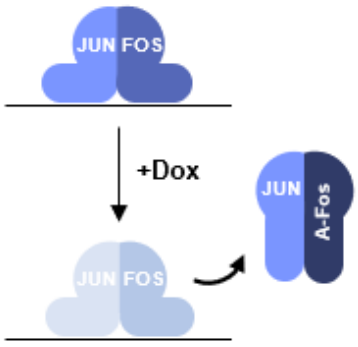


Figure 2- 29: Schematic showing the effect of A-FOS induction.

Upon doxycycline-induced overexpression of A-FOS, it dimerizes with Jun and thus prevents Fos:Jun heterodimer formation and chromatin binding.

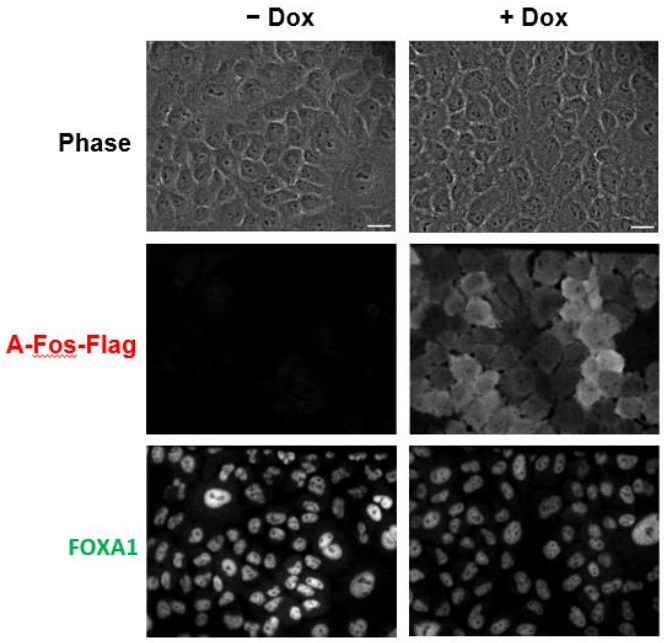


Figure 2- 30: Immunostaining of FLAG-tagged A-Fos and FOXA1 ± A-Fos induction.

Scale bar: 20 μm.

FOSL2 ChIP-seq verified that A-FOS expression leads to a near-complete inhibition of genome-wide AP-1 binding (**Figure 2-31&2-32**).

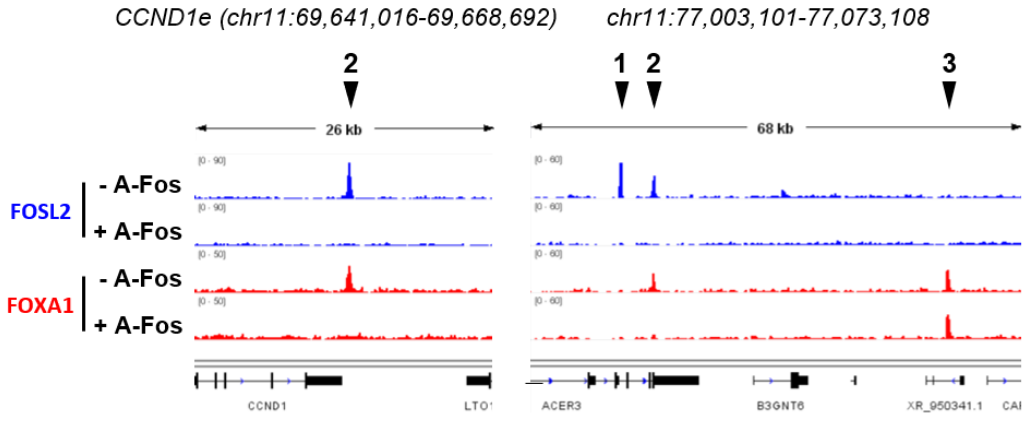


Figure 2- 31: Genomic tracks of FOSL2 and FOXA1 ChIP-seq ± A-Fos induction.

Arrows 1-3 demarcate examples of AP-1 unique, AP-1 / FOXA1 overlapped, and FOXA1 unique sites, respectively. In the presence of A-FOS, FOXA1 binding is significantly reduced at the overlapped site (2), but not the unique site (3).

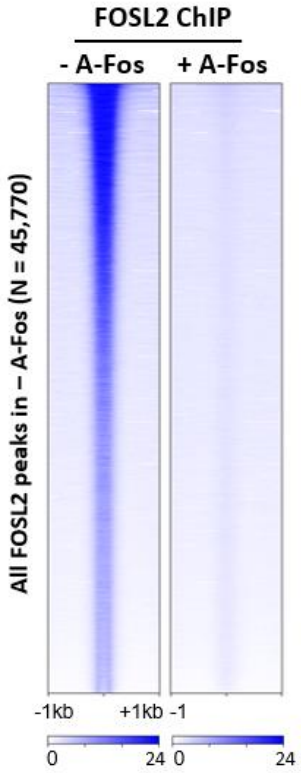


Figure 2- 32: Heatmaps showing FOSL2 ChIP-seq signal across genome-wide FOSL2 binding sites ± A-Fos induction.

FOXA1 ChIP-seq in the ±Dox conditions shows that A-FOS induction causes significant reduction of FOXA1 binding over the sites where FOXA1 and AP-1 peaks overlap, while FOXA1 binding over non-overlapping sites remains unchanged (**Figure 2-33&2-34**).

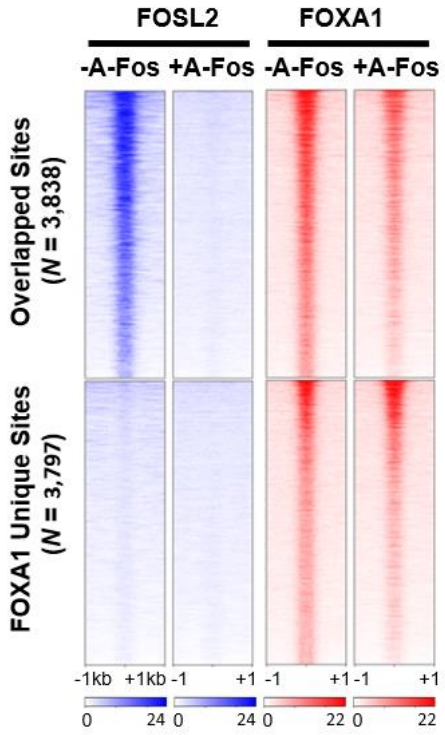


Figure 2- 33: Heatmap of FosL2 and FOXA1 ChIP-seq signals over AP-1 / FOXA1 overlapped sites and FOXA1 unique sites ± A-FOS.

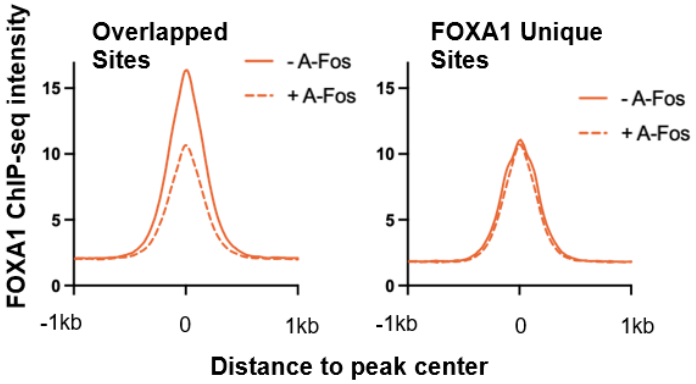


Figure 2- 34: Profiles of the average FOXA1 ChIP-seq intensities in Figure 2-33.

Differential binding analysis revealed 1,340 reduced and 234 enhanced FOXA1 peaks in the presence of A-FOS ("lost" vs "gained" peaks). Over 80% of the lost peaks contain AP-1

motifs and/or show AP-1 binding, much higher than the unchanged and the gained peaks (**Figure 2-35**). These results further support that AP-1 directs FOXA1 binding.

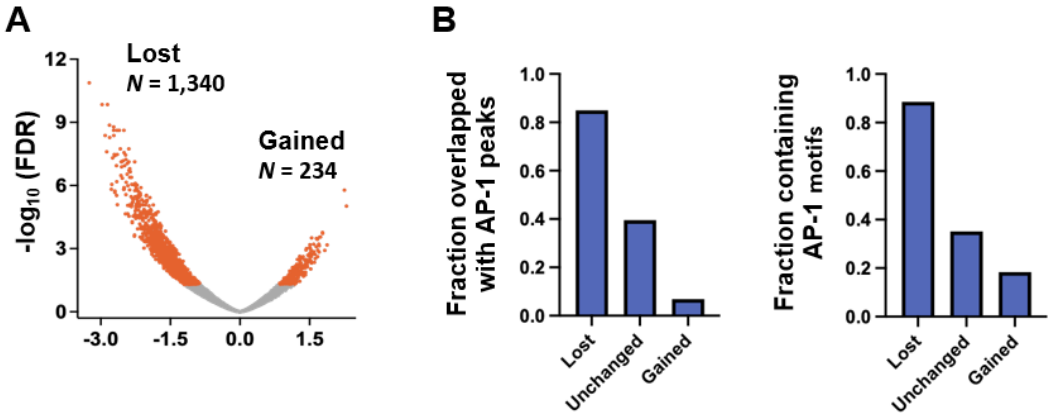


Figure 2- 35: differential FOXA1 binding ± A-FOS.

A) Volcano plot showing differential FOXA1 binding ± A-FOS. Reduced (loss) and enhanced (gain) FOXA1 peaks in +A-FOS are highlighted. B) Overlap of loss / unchanged / gain FOXA1 peaks with AP-1. The left panel shows the fraction that overlapped with FosL2 peaks, and the right panel shows the fraction that contains AP-1 motifs.

We next analyzed whether AP-1-enhanced FOXA1 binding depends on specific configurations of their motifs, i.e. relative orientation and distance (Figure 2-36).

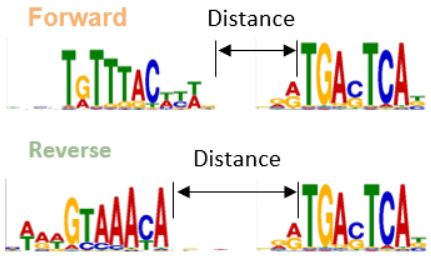


Figure 2- 36: Definition of the forward and reverse orientation between FOXA1 and AP-1 motifs, and the distance in between.

The two motif orientations are evenly distributed regardless of the peak category (**Figure 2-37**), while the two motifs are much more likely to be located within 8bp in the lost peaks (**Figure 2-38**). These results, along with the data in **Figure 2-20**, show that proximity, but not a

specific spacing or orientation between FOXA1 and AP-1 motifs, is required for their cooperativity. Weak enrichment is also observed near 10, 20, 30, and 40bp, suggesting that the rotational orientation of these two motifs on the same side of the DNA promotes cooperativity.

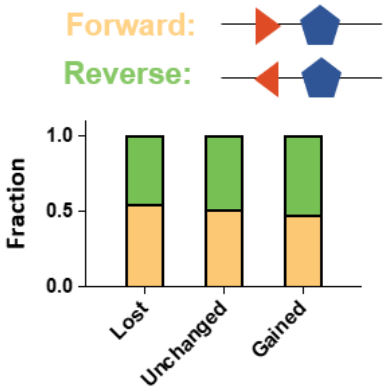


Figure 2- 37: Distribution of the two orientations of FOXA1 / AP-1 motifs in loss / unchanged / gain peaks.

Orange arrow: FOXA1 motif, blue pentagon: AP-1 (palindromic).

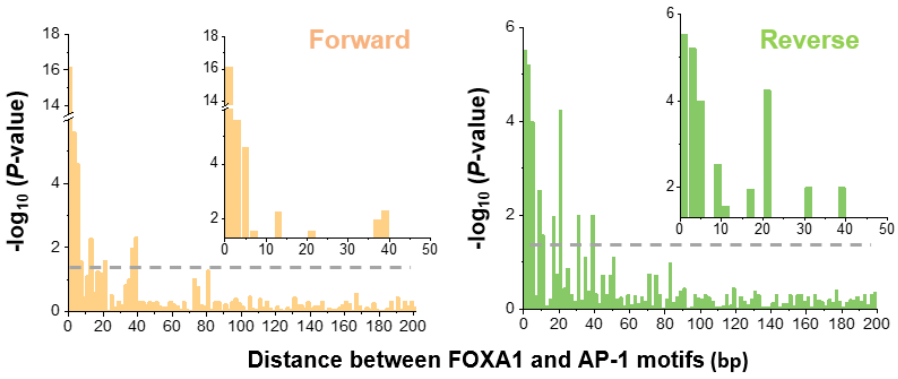


Figure 2- 38: Distribution enrichment of FOXA1 / AP-1 motif distances in loss peaks separated by the two orientations.

Enrichment was calculated based on the histogram using right-tailed two-proportion Z-test.

In addition, we found that the maximum FOXA1 motif scores are significantly lower in lost peaks than in unchanged and gained peaks (**Figure 2-39**). This is consistent with the observation in **Figure 2-24** that FOXA1 binding over weaker motifs tends to be more AP-1

dependent. It also suggests that, upon AP-1 inhibition, FOXA1 is released from the weaker sites and re-distributed to stronger motifs.

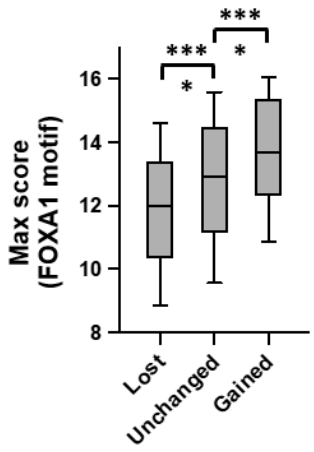


Figure 2- 39: The distributions of the maximum FOXA1 motif score per peak for loss / unchanged / gain FOXA1 peaks.

To explore the functional role of FOXA1 binding events potentiated by AP-1, we conducted RNA-seq in cells ±A-FOS overexpression. We found that the genes proximal to lost peaks show significant down-regulation in the absence of AP-1, while the ones associated with gained peaks tend to be upregulated (**Figure 2-40**). Differential expression analysis also revealed the same trend (**Figure 2-41**). These results indicate that AP-1-facilitated FOXA1 binding events mostly mediate positive regulation of gene expression in A549 cells.

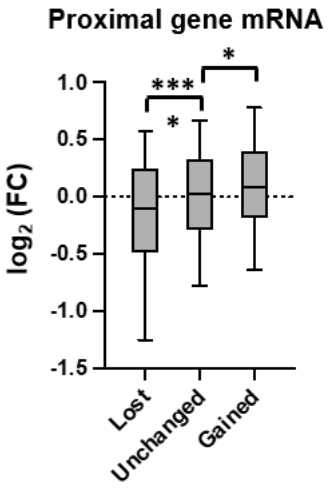


Figure 2- 40: Fold changes of the RNA-seq counts with A-FOS overexpression for the proximal genes near loss / unchanged / gain FOXA1 peaks.

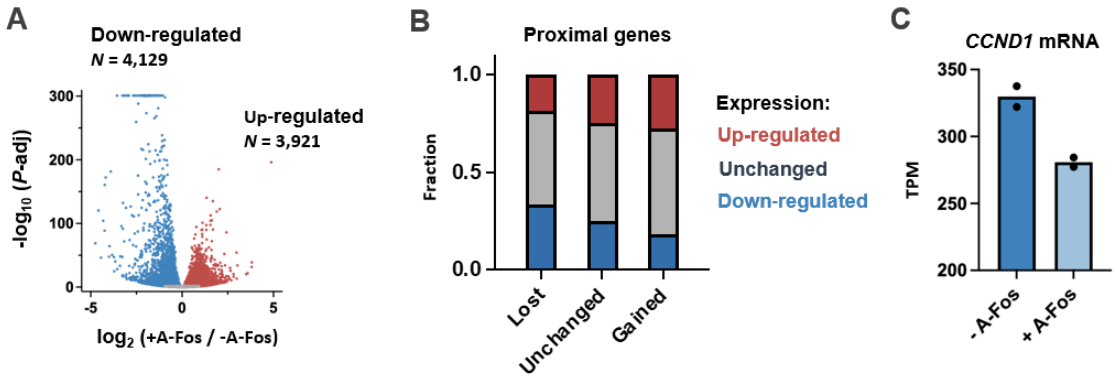


Figure 2- 41: Differential expression analysis of the RNAseq ± A-Fos induction.

A) Volcano plot of RNAseq counts in ± A-Fos conditions. Red and blue represent up-regulated vs down-regulated genes in the presence of A-Fos induction, respectively. B) Differential regulation of genes proximal to lost, unchanged, and gained FOXA1 peaks. The genes close to the lost peaks are more likely to be down-regulated. C) CCND1 mRNA is downregulated in the presence of A-Fos, consistent with the lost of FOXA1 binding in the CCND1 enhancer.

Gene ontology (GO) analysis of the genes proximal to the lost peaks show enrichment in the cell migration, tissue development, and signal transduction categories (**Figure 2-42**), implying their cell-type-specific and differentiation-linked functions.

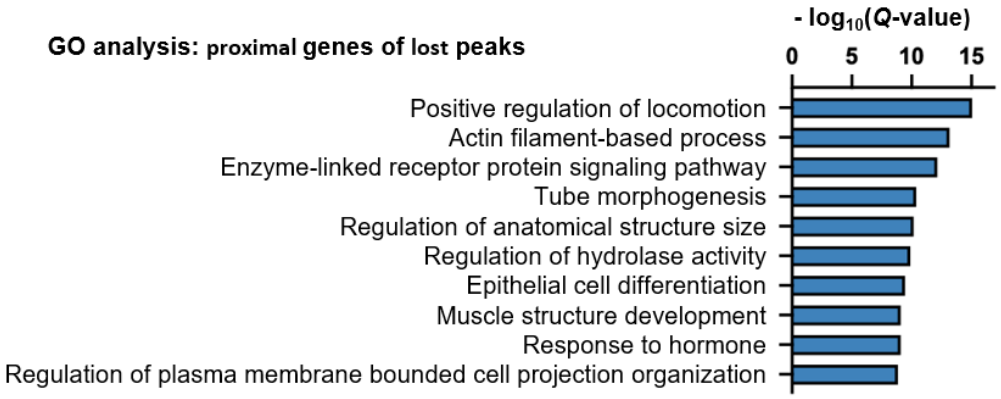


Figure 2- 42: Top 10 enriched Gene Ontology (GO) terms of the proximal genes near the loss FOXA1 peaks.

2.3.5 In vitro study of FOXA1 binding and cooperativity with AP-1.

To directly evaluate the intrinsic FOXA1 binding activity, we developed an electrophoretic mobility shift assay followed by sequencing (EMSA-seq) to measure the in vitro binding affinities between FOXA1 and all library sequences simultaneously (**Figure 2-43**). In this method, EMSA was performed using mixed library DNA incubated with purified FOXA1 at different concentrations (**Figure 2-44**). Shifted (FOXA1-bound) vs unshifted (unbound) bands were then purified, PCR amplified, and subjected to amplicon sequencing. Normalized sequencing counts were converted into the “ratio bound in vitro” for each sequence, which was highly correlated among two replicates.

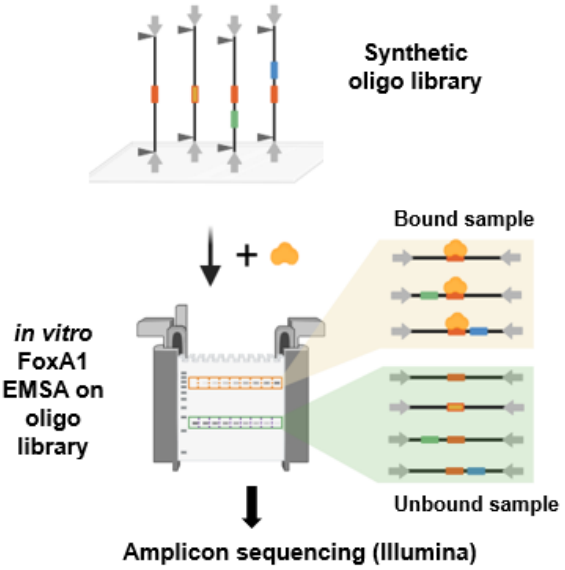


Figure 2- 43: Workflow of EMSA-seq.

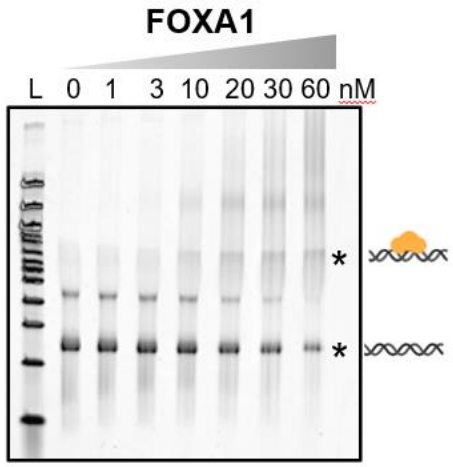


Figure 2- 44: A representative EMSA-seq gel conducted on the ISO library with increasing levels of purified recombinant mouse FOXA1.

The lower asterisk represents unbound oligonucleotides, and the upper asterisk indicates FOXA1-bound oligonucleotides. L: 100bp DNA ladder. Data collected by Holly Godin.

FOXA1 binding in vitro is primarily determined by the motif strength. Among the CCND1e variants, for example, FOXA1 binding generally increases with the number of motifs

(**Figure 2-45A**). The EMSA-seq signals of the whole library are highly correlated with FOXA1 motif score (**Figure 2-45B**). Different features of DNA shape play only a minor role (**Figure 2-46**).

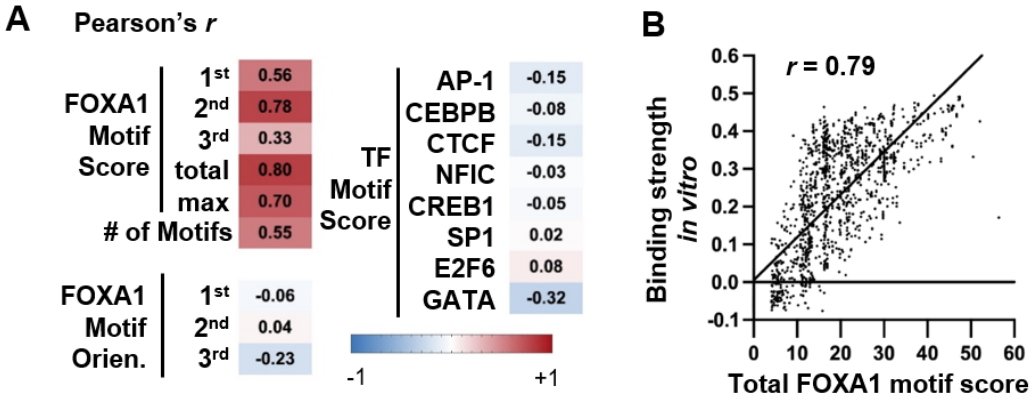


Figure 2- 45: Pearson correlation coefficient between FOXA1 binding strength in vitro and CCND1 sequence variables.

A) Pearson correlation coefficient. B) Correlation between the binding strength in vitro and total (summed) FOXA1 motif score for each sequence. Data collected by Holly Godin.

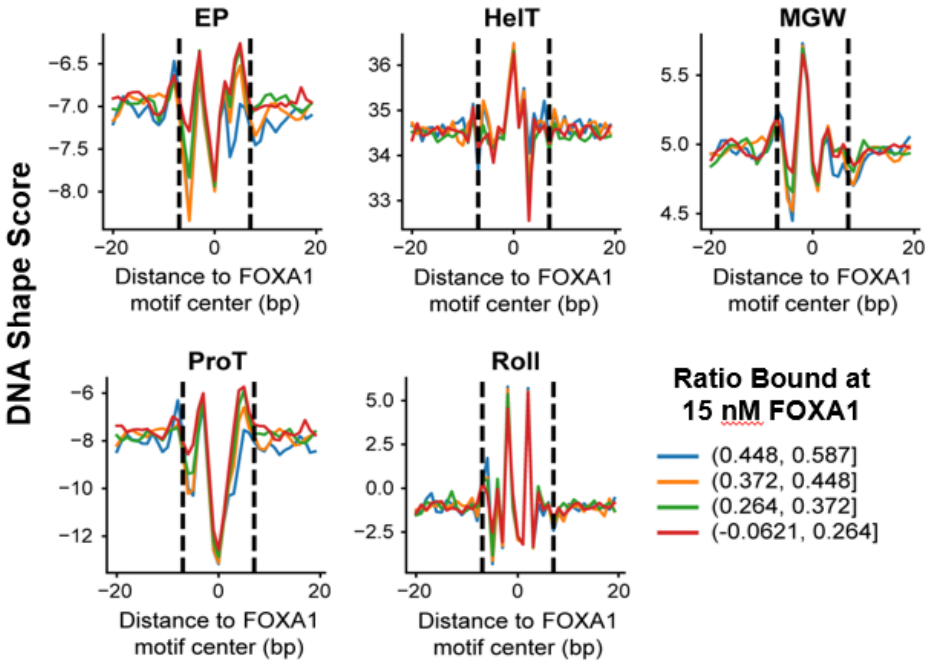


Figure 2- 46: Plots showing five DNA shape parameters averaged across four sets of FOXA1 binding sites.

DNA shape parameters include electrostatic potential (EP), helix twist (HelT), minor groove width (MGW), propeller twist (ProT), and roll. Four colors represent different range of in vitro binding strength. Data collected by Jianyu Yang.

Importantly, FOXA1 binding is no longer sensitive to mutations in AP-1 motifs in vitro (**Figure 2-45A&2-47A**), confirming that the AP-1 effect on the FOXA1 ChIP-ISO signals is not due to inadvertent changes in intrinsic FOXA1 binding affinities. On the same library sequences, in vitro FOXA1 binding poorly correlates with that in vivo (**Figure 2-47B**). This reinforces our previous finding that strong FOXA1 motifs are only partially responsible for FOXA1 binding in vivo.

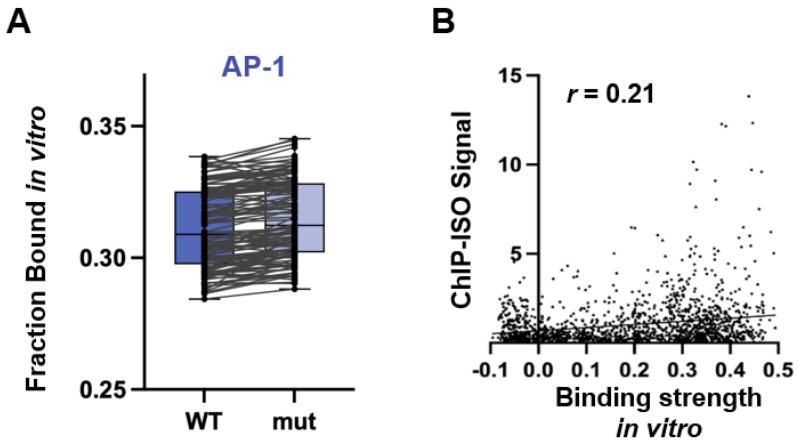


Figure 2- 47: Box-and-whisker plot and Correlation between the binding strength in vivo (measured by ChIP-ISO) and that in vitro (EMSA-seq).

A) Box-and-whisker plots showing the changes of FOXA1 EMSA-seq signal on templates ±AP-1 motif. B) Correlation between in vivo and in vitro. Data collected by Holly Godin.

To investigate potential cooperativity between AP-1 and FOXA1 in vitro, we purified recombinant AP-1 and performed low-throughput EMSAs with AP-1 and FOXA1 using CCND1e DNA. To focus on the co-binding between AP-1 and the most proximal FOXA1 motif, as indicated by **Figure 2-13**, we used a CCND1e template that has the first two FOXA1 motifs mutated (**Figure 2-48A**). The gels show distinct bands for DNA bound by FOXA1 or AP-1 alone, and a super-shift for DNA bound by both factors (**Figure 2-48A**). Quantification of the unbound band intensity shows that the presence of AP-1 moderately promotes the binding of FOXA1, and vice versa (**Figure 2-48B**).

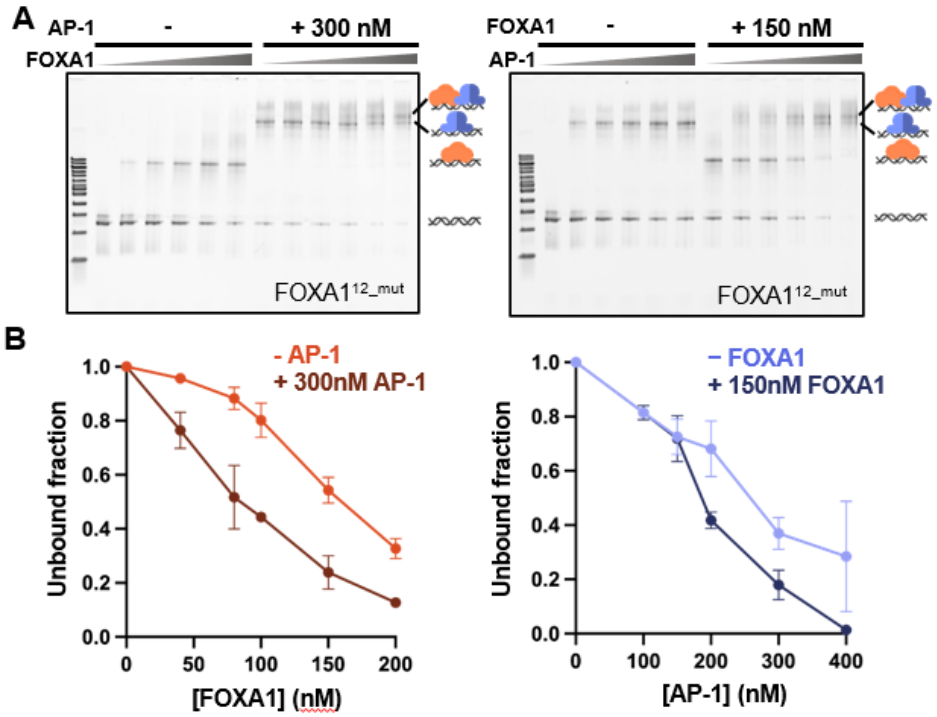


Figure 2- 48: Cooperative binding of FOXA1 and AP-1 *in vitro* on naked DNA template of *CCND1e*.

A) Representative EMSA gels with FOXA1 titration ± AP-1 (left) or AP-1 titration ± FOXA1 (right) using a CCND1 variant with the first two FOXA1 motifs mutated (FOXA112_mut). Different populations are labeled on the right side of the gel (FOXA1 = orange, AP-1 = blue). B) Quantification of the EMSA gel in panel A. Error bar represents standard error for three replicates. Data collected by Holly Godin.

Interestingly, we observed binding cooperativity between FOXA1 and AP-1 even in the absence of one factor's motif (**Figure 2-49**). These results suggest that FOXA1 and AP-1 may exhibit protein-protein interactions that allow them to recruit each other without direct DNA binding. This can at least partially explain the interdependency and cooperativity of these two factors in vivo.

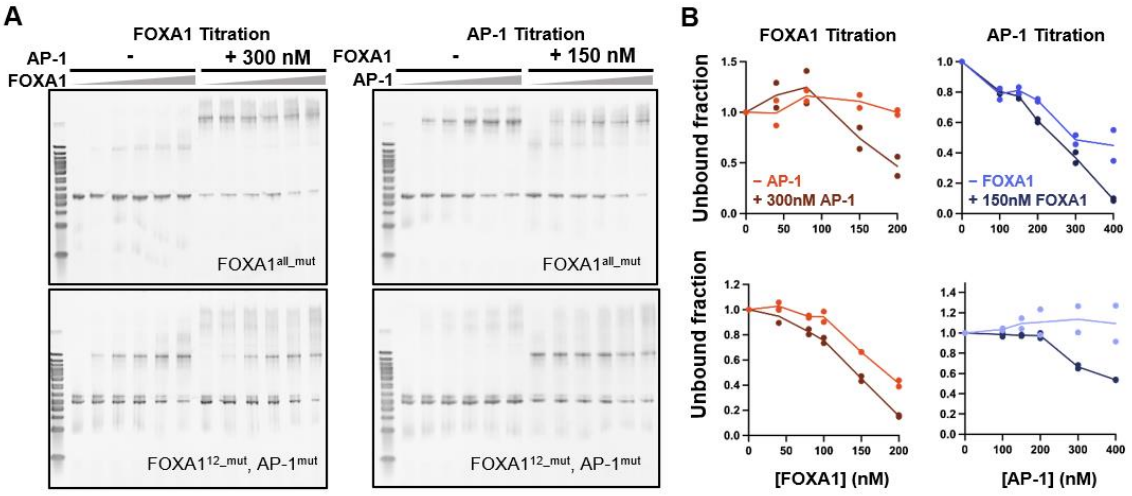


Figure 2- 49: Same as Figure 2-48 but using different DNA templates and with two replicates performed on each template.

Data collected by Holly Godin.

2.3.6 FOXA1 binding is mostly determined by the local sequence, not the chromatin context.

With the work above focusing on local sequences, we next explored how the larger-scale chromatin context can impact FOXA1 binding. In ChIP-ISO, native sequences containing FOXA1 motifs are moved from their endogenous loci to the euchromatic AAVS1 site (**Figure 2-50**). Comparison of FOXA1 occupancy at the native vs AAVS1 locus therefore allows us to infer the effect from the endogenous chromatin (**Figure 2-51**).

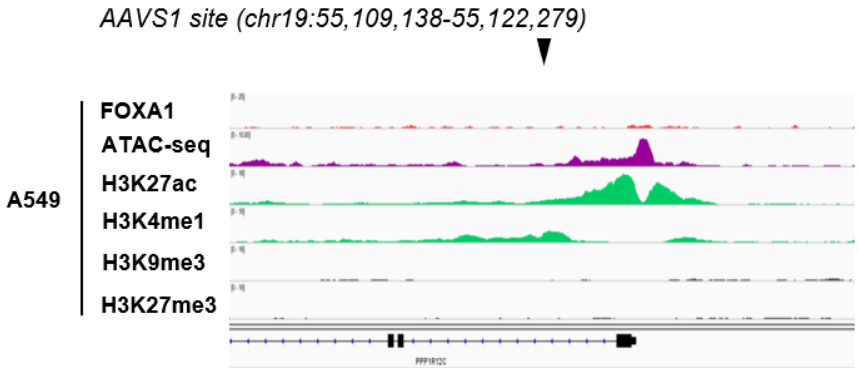


Figure 2- 50: Genomic tracks of TF and histone modification ChIP-seq and ATAC-seq signals at the AAVS1 safe-harbor locus in WT A549 cells.

Arrow indicates locus where library sequences were inserted. The insertion site is near active histone marks, but not the repressive ones.

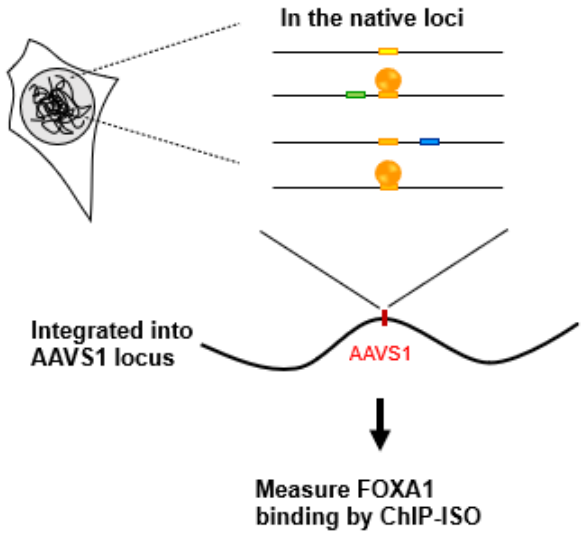


Figure 2- 51: Strategy for studying the effect of chromatin context.

We first applied this strategy to FOXA1 sites within euchromatic regions. We selected two sets of native sequences where FOXA1 binding cannot be explained by its motif strength: those with high-score FOXA1 motifs but mostly weak binding (set one) and vice versa (set two) (**Figure 2-52&2-53**). EMSA-seq shows that most of these sequences are bound by FOXA1 in

vitro, with slightly higher occupancies in set one (**Figure 2-53**). Strikingly, ChIP-ISO signals on these sequences at the AAVS1 locus largely recapitulate the ChIP-seq intensities at the endogenous sites (**Figure 2-53**), indicating that FOXA1 binding are mostly determined by the local sequences. Such local signals again involve co-factors, with set two sites being more enriched with AP-1 and CEBPB motifs and showing higher AP-1 and CEBPB binding (**Figure 2-53**).

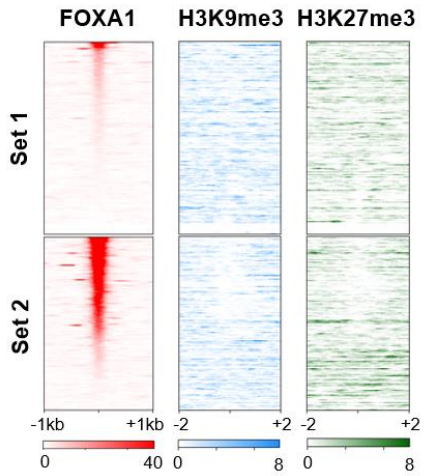


Figure 2- 52: ChIP-seq signals of repressive histone marks near the native sequences included in Figure 2-53.

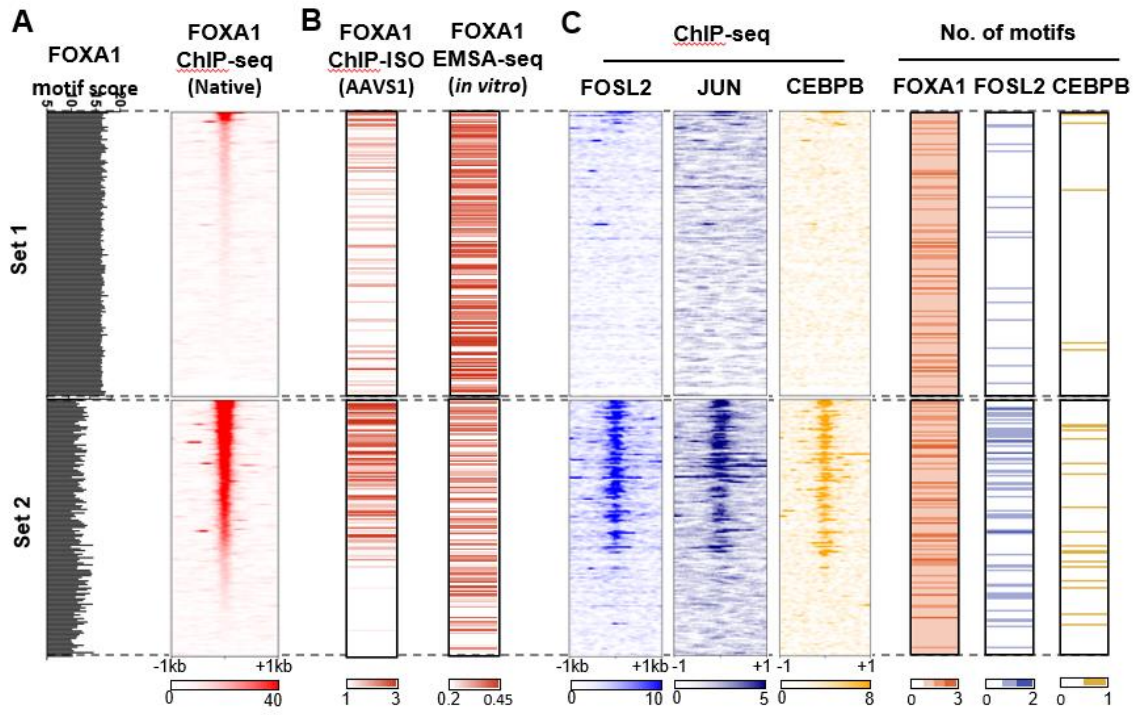


Figure 2- 53: FOXA1 binding in euchromatin is mostly determined by the local sequence, not the chromatin context.

A) ChIP-ISO test cases where strong / weak FOXA1 motifs show low / high FOXA1 binding (set 1 and 2, respectively). Left: FOXA1 motif scores. Right: heatmap of FOXA1 ChIP-seq signals in WT A549. B) Heatmap of FOXA1 ChIP-ISO signals (left) and EMSA-seq signals (fraction bound at 15 nM FOXA1, right) for sequences in panel A. C) Heatmap of FOSL2, JUN, and CEBPB ChIP-seq signals in WT A549 (left) and number of FOXA1, FOSL2, and CEBPB motifs for sequences in panel A.

We next investigated the effect of heterochromatin by selecting FOXA1 motifs from regions covered by H3K9me3 and H3K27me3. If heterochromatin has a strong inhibitory effect, we expect FOXA1 binding to increase when these sequences are transferred to euchromatin (**Figure 2-54**).

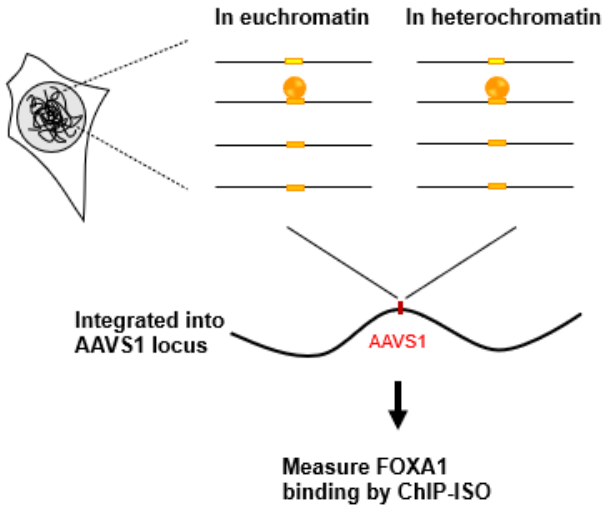


Figure 2- 54: Strategy for studying the effect of heterochromatin.

We picked 56 sequences from H3K9me3-marked regions (**Figure 2-55A**). For comparison, we assembled a control set in euchromatin that lacks H3K9me3 but exhibits matching levels of FOXA1 binding (**Figure 2-55**). The differences in H3K9me3 signals between these two sets of regions disappear after they are relocated to the AAVS1 locus, confirming the elimination of H3K9me3 marks at the new site (**Figure 2-55B**). However, this does not lead to enhanced FOXA1 binding, as the sequences originally from heterochromatin still exhibit the same FOXA1 binding levels as the euchromatic control (**Figure 2-55C**).

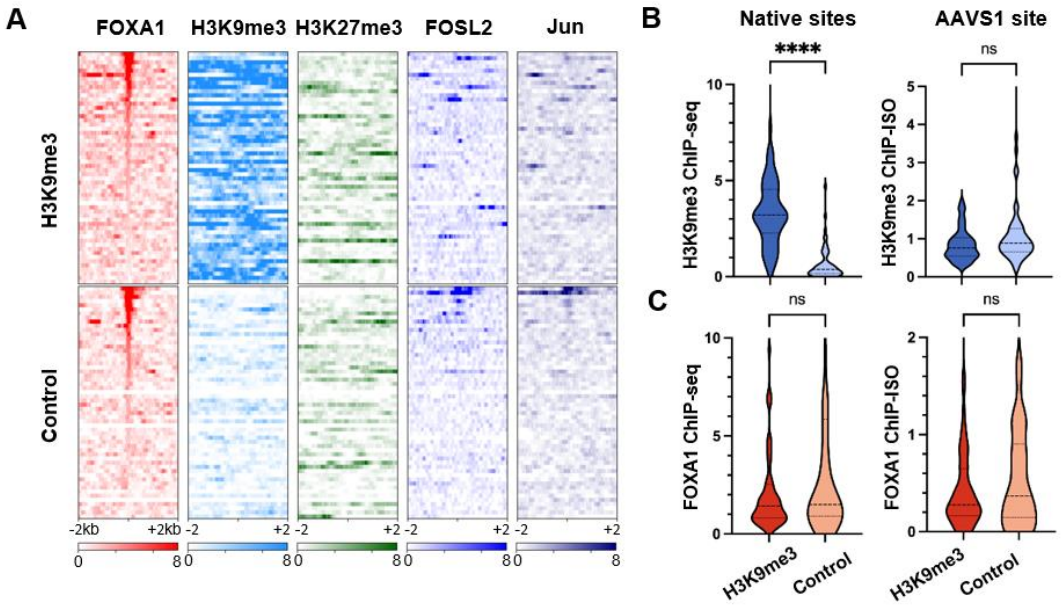


Figure 2- 55: FOXA1 binding in H3K9me3-marked heterochromatin is mostly determined by the local sequence, not the chromatin context.

A) Heatmap of FOXA1, H3K9me3, H3K27me3, FOSL2, and Jun ChIP-seq signals in WT A549 for a set of ChIP-ISO library sequences derived from H3K9me3-marked regions (top) and a control set with comparable FOXA1 binding derived from euchromatic loci (bottom). B) Violin plot of H3K9me3 ChIP-seq signals for sequences in panel d at their native genomic loci (left) and ChIP-ISO signals of the same sequences at AAVS1 (right). C) Same as panel B, but for FOXA1. Data collected by Holly Godin.

We performed the same experiments using FOXA1 sites from H3K27me3 regions and got similar results (**Figure 2-56**). Combining the data from eu- and heterochromatin, FOXA1 binding at the AAVS1 site is highly correlated with that in their native sites (r = 0.69) (**Figure 2-57**). Overall, this data suggests that the native chromatin context, including H3K9me3 and H3K27me3 marks, plays a minor role in FOXA1 binding.

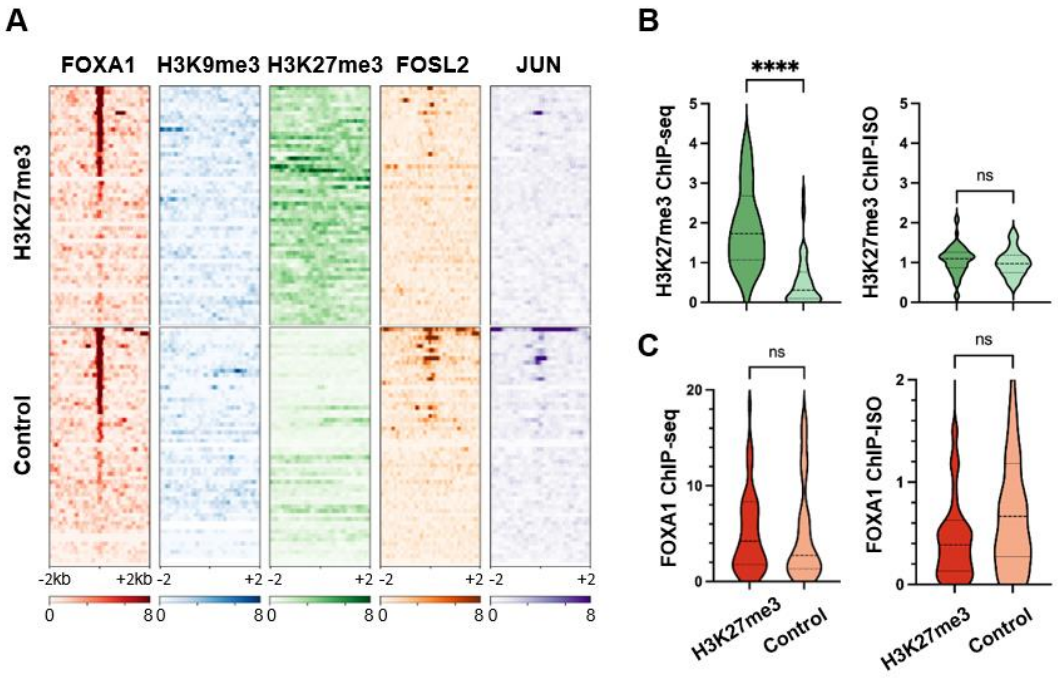


Figure 2- 56: FOXA1 binding in H3K27me3-marked heterochromatin is mostly determined by the local sequence, not the chromatin context.

A) Heatmap of FOXA1, H3K9me3, H3K27me3, FOSL2, and Jun ChIP-seq signals for a set of ChIP-ISO library sequences derived from H3K27me3-marked genomic loci (top) and a set of control sequences with comparable FOXA1 binding level derived from euchromatic loci (bottom). B) Violin plot of H3K27me3 ChIP-seq signals for sequences in panel A at their native genomic loci (left) and ChIP-ISO signals of the same sequences at AAVS1 (right). Dark green: H3K27me3 set, and light green: control set from panel A. ****: $p < 0.0001$ and ns: non-significant. C) Same as panel B, but for FOXA1. Data collected by Holly Godin.

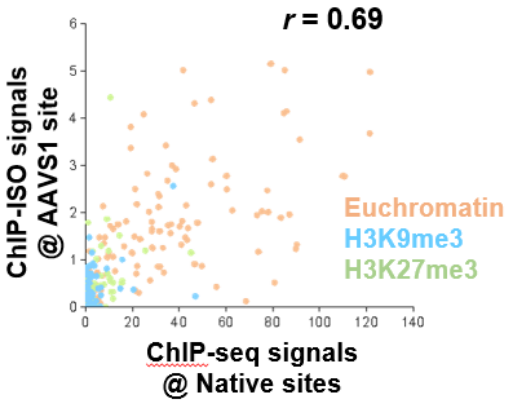


Figure 2- 57: Correlation between FOXA1 ChIP-ISO and ChIP-seq signals for all sequences derived from the native genome in our library.

Orange: sequences from euchromatin, blue: H3K9me3-marked heterochromatin, green: H3K27me3-marked heterochromatin.

To test whether H3K9me3 enables better prediction of FOXA1 binding genome-wide, we again turned to neural networks trained on FOXA1 ChIP-seq data. Specifically, we used our previously described Bichrom neural network architecture to integrate DNA sequence and various chromatin features into the training process[171]. Integrating ATAC-seq signals or a combination of histone marks into the neural network helps to improve performance in distinguishing held-out FOXA1-bound and unbound sites (**Figure 2-58**). However, integrating H3K9me3 alone alongside DNA-sequence features does not improve performance (**Figure 2-58**), suggesting that Bichrom is unable to learn any informative relationship between H3K9me3 and FOXA1 binding.

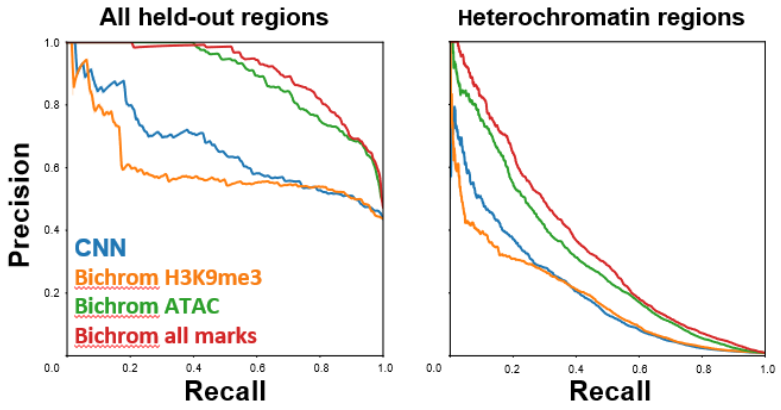


Figure 2- 58: Precision recall curves showing performance of neural networks trained on FOXA1 ChIP-seq data in A549 cells.

Each plot shows performance of a CNN trained using only sequence (blue lines); Bichrom trained using sequence and H3K9me3 (orange lines); Bichrom trained using sequence and ATAC-seq (green lines); and Bichrom trained using sequence and a selection of five histone marks (red lines). The left plot shows the performance of the neural networks across all held-out test sites, while the right plot shows performance at FOXA1 motif instances that overlap H3K9me3 or H3K27me3 peaks. Data collected by Jianyu Yang.

We noted that FOXA1 motifs in H3K9me3 covered regions are weaker and have no adjacent AP-1 motifs. To ensure that the absence of a heterochromatin effect is not simply due to

the lack of suitable motifs, we artificially introduced H3K9me3 to a strong FOXA1 site by targeting KRAB-dCas9 to the endogenous CCND1e through CRISPRi (**Figure 2-59A**). ChIP-qPCR shows robust H3K9me3 signal at the CCND1e upon 48 hours of KRAB-dCas9 induction compared to uninduced cells (**Figure 2-59B**). Upon H3K9me3 deposition, we found that FOXA1 binding at the CCND1e does not significantly change compared to uninduced cells (**Figure 2-59B**). These results are in line with our finding that native chromatin context plays only a minor role, if any, in FOXA1 binding. Together, we conclude that FOXA1's binding specificity in vivo is more determined by the local sequence than the epigenetic background.

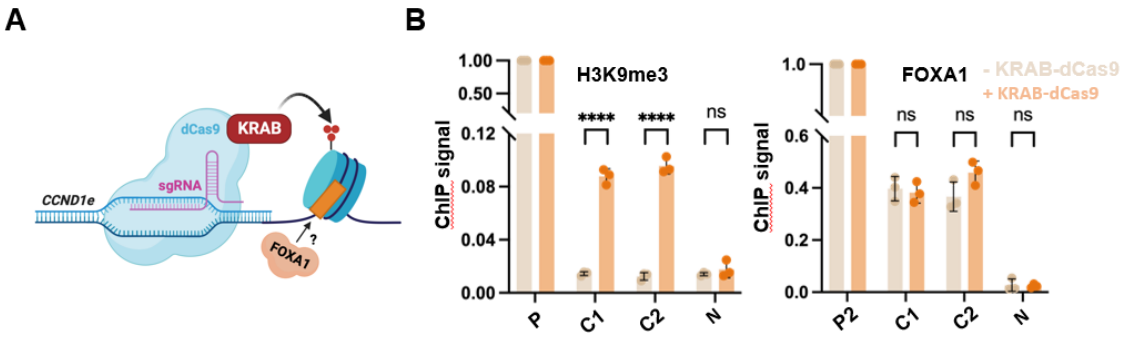


Figure 2- 59: H3K9me3-marked heterochromatin does not directly repress FOXA1 binding at *CCND1e*.

A) Schematic of CRISPRi method where KRAB-dCas9 is induced by doxycycline to ectopically write H3K9me3 to the endogenous CCND1e in A549 cells. This system is used to measure the effect of H3K9me3 deposition on FOXA1 binding. B) H3K9me3 (left) and FOXA1 (right) ChIP-qPCR signals on three biological replicates at a positive control region (P), CCND1e (C1 and C2), and a negative control region (N) with (orange) or without (beige) KRAB-dCas9 induction. ****: $p < 0.0001$ and ns: non-significant based on two-tailed unpaired t-test. Data collected by Holly Godin.

2.3.7 Cell-type-specific binding of FOXA1 correlates with differential expression of AP-1.

Previous studies have shown that FOXA1 has different binding patterns in different cell types[115]. Considering our findings above, we hypothesized that differential availability of co-factors may contribute to such cell-type specificity. We therefore analyzed the RNA-seq data of

FOXA1 and AP-1 subunits in three cancer cell lines, A549, HepG2, and MCF-7. Interestingly, FOXA1 mRNA level is lower but AP-1 subunits are higher in A549 than the other two cell lines (**Figure 2-60**). Immunostaining confirmed this trend at the protein level (**Figure 2-61**).

TPM		A549	HepG2	MCF-7
FOXA1		13.7	34.7	169.4
FOS	FOS	173.3	16.1	56.5
	FOSB	29.4	2.1	6.3
	FOSL1	6.2	2.6	1.3
	FOSL2	36.8	14.6	28.4
JUN	JUN	35.4	17.7	10.4
	JUNB	64.0	9.2	26.3
	JUND	46.7	70.1	22.2

Figure 2- 60: RNA-seq counts, reported in transcripts per million (TPM), for FOXA1 and various AP-1 subunits in WT A549, HepG2, and MCF-7 cell lines.

Red and blue mark the highest and lowest expression for each gene.

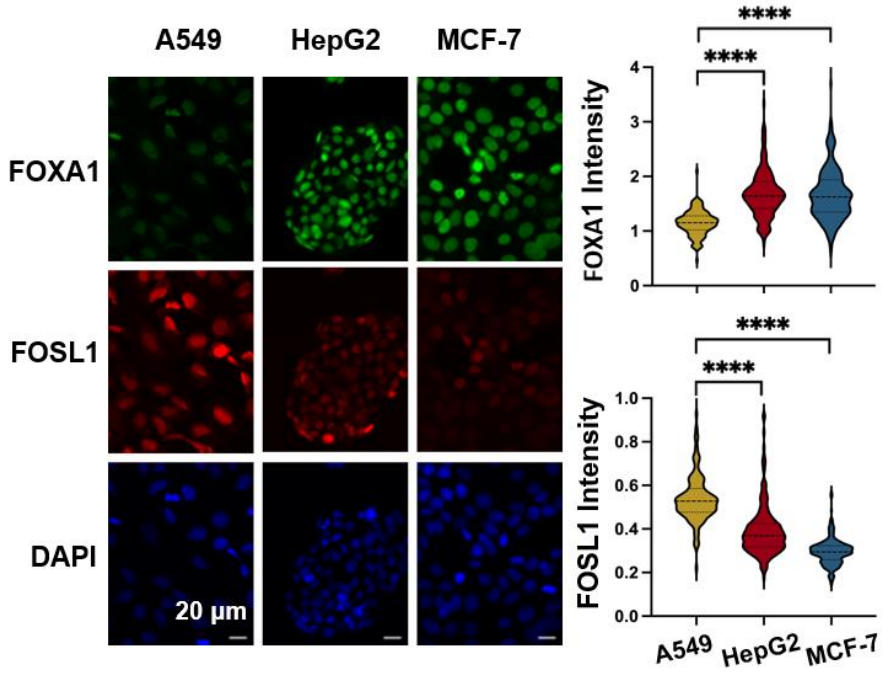


Figure 2- 61: Differential expression of FOXA1 and FOSL1 in A549, HepG2 and MCF-7 cells.

Immunostaining of FOXA1 (green), FOSL1 (red), and nuclei (DAPI) in fixed A549, HepG2, and MCF-7 cells. Right: Violin plots showing normalized FOXA1 (top) and FOSL1 (bottom) intensities in these three cell lines. Scale bar: 20 μm.

Despite the lower expression of AP-1 in HepG2 and MCF-7, FOXA1 still co-binds with AP-1, but the level of overlap is significantly reduced compared to A549 (**Figure 2-62**).

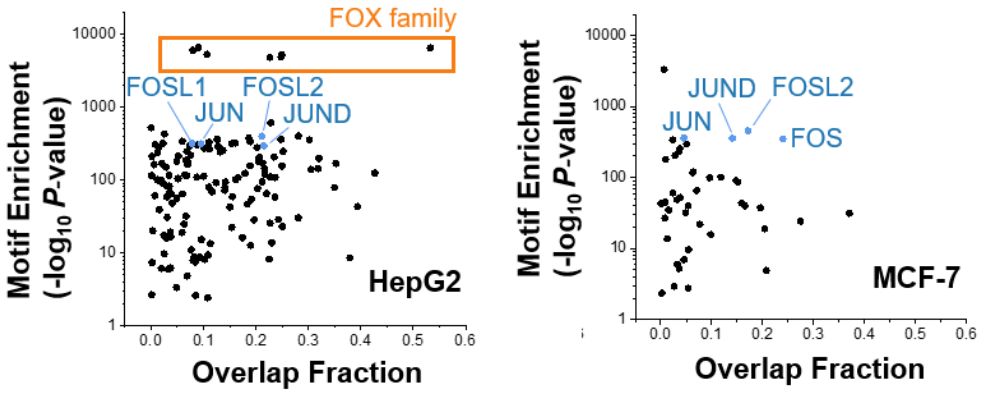


Figure 2- 62: Bioinformatic analysis of FOXA1 / TF co-binding in HepG2 (left) and MCF-7 (right).

For each TF, the dot plot shows the percentage of the overlapped FOXA1 ChIP-seq peaks (x axis) and the enrichment of its motif within FOXA1 peaks (y axis). AP-1 subunits are indicated in blue. Data collected by Holly Godin.

The observations above raise the possibility that the abundance of AP-1 in A549 allows it to play a more dominant role in directing FOXA1 binding than in HepG2 and MCF-7. If this is true, we would expect a higher representation of AP-1 motifs in A549-specific FOXA1 peaks than the HepG2 or MCF-7-specific peaks. To test this prediction, we performed differential binding analysis of FOXA1 in A549 vs MCF-7/HepG2 and searched for motifs in common and cell-type specific peaks (**Figure 2-63A**). Indeed, A549-specific FOXA1 peaks are much more likely to contain AP-1 motifs, compared with shared peaks or MCF-7/HepG2-specific peaks (**Figure 2-63B**). In A549 cells, AP-1 binding strongly correlates with that of FOXA1, and such correlation is weaker in the other two cell types (**Figure 2-63A**). Mild AP-1 binding, however, is still detectable over MCF-7 and HepG2-specific FOXA1 sites, despite the lower enrichment of

AP-1 motifs. This may be due to the tethering of AP-1 by FOXA1 as indicated by the in vitro binding assay (**Figure 2-49**). Overall, this data indicates that AP-1 is at least partially responsible for cell-type-specific binding of FOXA1.

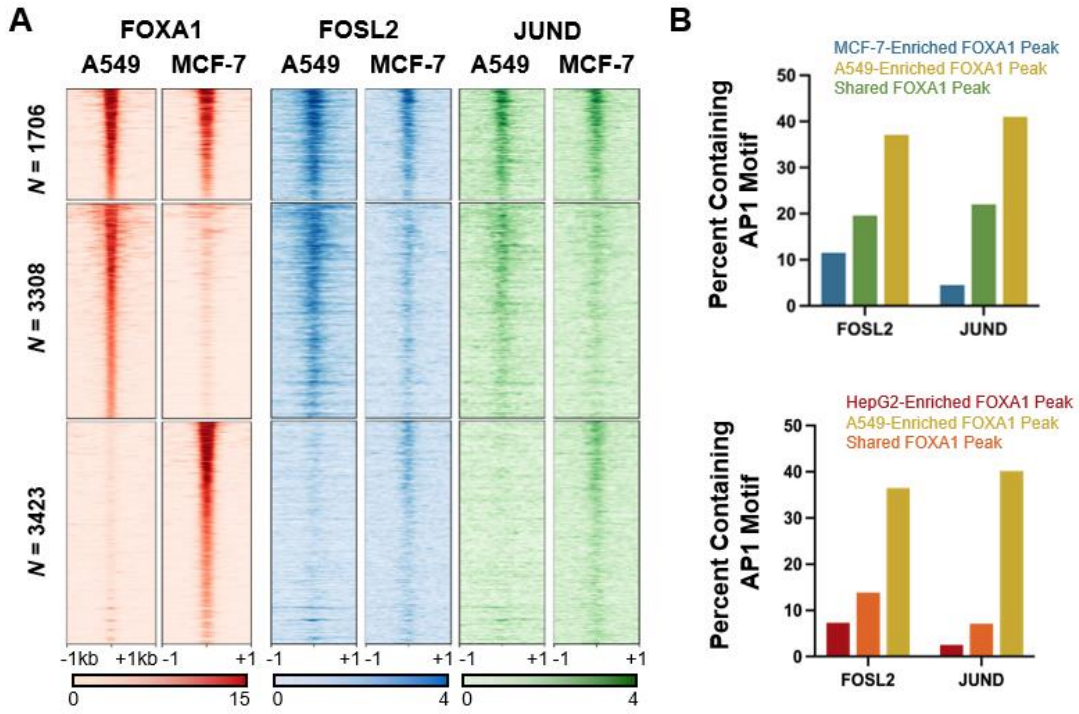


Figure 2- 63: Differential FOXA1 binding analysis by occurrence probability of FOSL2 or JUND motif in common or differential FOXA1 peaks.

A) Differential FOXA1 binding analysis in A549 and MCF-7 cells, with heatmaps showing common (upper), A549-specific (middle), and MCF-7-specific (bottom) peaks. FOSL2 and JUND ChIP-seq signals over the same regions are shown on the right. B) Occurrence probability of FOSL2 or JUND motif in common or differential FOXA1 peaks. Upper: A549 vs MCF-7; Lower: A549 vs HepG2. Data collected by Holly Godin.

To further characterize the importance of AP-1 in specifying cell-type-specific FOXA1 binding, we trained DNA sequence neural networks on 13 cell lines and tissue types, including the three above, with published FOXA1 ChIP-seq data. We again performed feature attribution analysis at ChIP-seq peaks and compiled informative patterns into motifs using TF-MoDISco. Consistent with the analysis above, AP-1 and CEBPB are not top features for promoting FOXA1 binding in MCF-7 and HepG2 (**Figure 2-64**). Interestingly, the AP-1 motif is identified as a

highly informative feature for FOXA1 binding in the RT4 urinary bladder cell line, suggesting that AP-1 may assist FOXA1 binding in different cell types. In other cell lines, our neural networks identify additional informative co-factor motifs, including CTCF in 22Rv1 prostate carcinoma epithelial cells and AP-2 in both GP5d and SK-BR-3 cancer cell lines. We speculate that these factors may play analogous roles to AP-1 in assisting FOXA1 binding in these cell types.

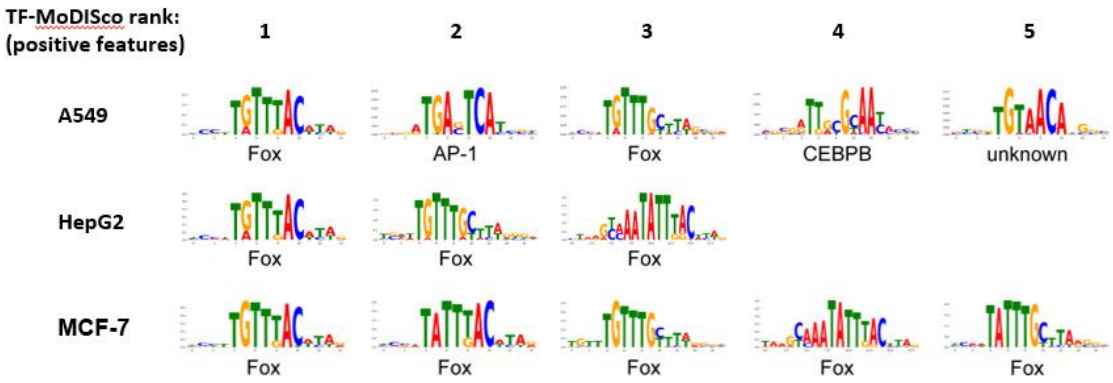


Figure 2- 64: Top ranking motifs detected by TF-MoDISco in the genome-wide DeepLIFT-Shap positive feature attribution scores at predicted sites.

The top five ranking motifs are shown unless TF-MoDISco returned fewer than five motifs. The TF family of best-matching motifs is annotated for each motif. Data collected by Jianyu Yang.

2.4 Discussion

Although PFs possess the unique ability to target nucleosomal motifs, they only bind a small subset of their motifs in higher eukaryotic genomes. Despite recent advances in genomic technology, extracting the genetic rules that govern TF/PF binding remains a formidable challenge. A significant hurdle arises from the fact that many genetic and epigenetic features can influence TF binding, and native genomes do not provide sufficient diversity to explore all possible combinations of these variables, especially within the constraints of evolution. In

comparison, the ChIP-ISO assay utilizes artificially designed sequences[172] that can circumvent evolutionary constraints and systematically perturb one genetic feature at a time. The synthetic sequences are inserted into the same genomic locus, which eliminates many variables caused by chromatin background, including the well-known ChIP artifact found near highly expressed genes[45173]. Overall, the ChIP-ISO assay allows us to quantitatively dissect the contribution to TF binding from individual genetic features.

Cooperative binding is a commonly reported phenomenon among many TFs in mammalian cells, but PFs are often thought to function independently as the most upstream factors that interact with chromatin. FOXA1, for example, was shown to be able to associate with a high affinity motif embedded in a compact nucleosome array and open local chromatin in vitro without the assistance of other TFs or remodelers[118]. It is therefore surprising that co-factors are the main determinant of FOXA1 binding in A549 cells. These seemingly contradictory findings may be reconciled by considering the FOXA1 motif strength: while FOXA1 is able to bind to a subset of strong motifs in the absence of co-factors (**Figure 2-24**), its binding on sub-optimal sites is strongly promoted by AP-1 and/or CEBPB (**Figure 2-21&2-39**). Co-binding with other TFs provides a mechanism for context-specific PF binding. Indeed, reduced AP-1 expression in MCF-7 and HepG2 cells releases FOXA1 near AP-1 motifs and allows it to occupy other genomic loci (**Figure 2-63**). This agrees with previous findings that the genomic distribution of FOXA1/2 can be affected by steroid receptors[119], GATA4[114], or PDX1[120]. These co-factor-dependent weak sites tend to be situated in open chromatin with active histone marks and are therefore more likely to carry regulatory functions (**Figure 2-24**). Consistently, native enhancers often contain sub-optimal motifs with reduced TF binding affinities[174,175]. Overall, these findings suggest that, although FOXA1 may use consensus motifs to engage with chromatin by itself, weaker motifs in conjunction with co-factors may play a more functional role during development to generate cell-type-specific binding and regulation[120,176].

AP-1 is a ubiquitously expressed TF that is highly represented in distal enhancers in many cell types[17751], where multiple TFs bind in hubs. It is therefore not surprising that the AP-1 motif is enriched near the binding sites of many TFs[177]. More relevant to this work, AP-1 was shown to co-bind with FOXA1 in breast and prostate cancer cells[115,178,179] and with FOXA1/2 in pancreatic ductal adenocarcinoma[180]. In most of these cases, however, it is not clear if these factors bind independently, cooperatively, or hierarchically. Here, using well-controlled synthetic sequences, we clearly demonstrated that the binding of FOXA1 and AP-1 is mutually dependent. This finding, together with a previous proposal that AP-1 may also function as a PF[181], raises an intriguing possibility that FOXA1 and AP-1 may bind together to achieve sufficient pioneering activities for the invasion and remodeling of closed chromatin in A549, which may contribute to cell-type-specific enhancer selection[182].

The molecular mechanism of FOXA1 and AP-1 binding cooperativity requires further elucidation, but our current data provides some clues. First, the in vitro EMSA assay shows that AP-1 enhances FOXA1 binding, and vice versa. Such effects can even be observed on templates that lack the motif for one of the factors. These results indicate that there may be protein-protein interactions between these two factors that allow one to be tethered by the other. Consistent with this idea, weak AP-1 binding can be detected near many MCF-7 or HepG2 enriched FOXA1 binding sites, despite the fact that >90% of these sites lack the AP-1 motif (**Figure 2-63**). Second, AP-1 can stimulate FOXA1 binding with different distances and orientations between their motifs (**Figure 2-37&38**). This argues against a model where FOXA1 and AP-1 form a rigid complex with highly specific interactions. It is more likely that they have DNA-dependent weak and polymorphic interactions, and this type of "soft motif syntax" is commonly found among cooperative TFs[66,183,184]. Third, although the cooperativity between FOXA1 and AP-1 is detected in vitro on naked DNA, the effect is weaker than that in vivo. It is therefore possible that their

cooperativity in vivo is also promoted by nucleosomes[185]. This would be consistent with the idea that FOXA1 and AP-1 may have co-pioneering activities.

Finally, our study suggests that the genomic background, including the heterochromatin marks, plays a minor role in FOXA1 binding. Heterochromatin has been reported in literature to both permit and inhibit PF binding[80,116,186]. For example, one study overexpressed a FOXA1 homolog, FOXA2, in immortalized foreskin fibroblasts cells and found that FOXA2 enrichment was generally depleted in H3K9me3 and H3K27me3 regions[114]. However, the same study also pointed out that, instead of a direct inhibitory effect from heterochromatin, this may be due to the lack of FOXA2 binding sites within these domains. Another study found that gain of FOXA2 binding at lamin-enriched sites is correlated with loss of H3K9me3, indicating its ability to target heterochromatin[187]. Here, we used two orthogonal methods, coupled with neural networks, to better understand the causal effect of heterochromatin on FOXA1 binding. Our ChIP-ISO results revealed that FOXA1 binds similarly to sequences in euchromatic vs heterochromatic context. Introducing ectopic H3K9me3 modifications over a strong FOXA1 site had no detectable effect on its binding. In addition, incorporation of H3K9me3 in the neural network analyses does not improve FOXA1 binding prediction. Together, these results suggest native chromatin context at most plays a minor role in FOXA1 binding. In summary, using the novel ChIP-ISO approach, in combination with in vitro and in silico analyses, our study demonstrated that cooperative binding with co-factors is the primary mechanism by which PFs achieve binding specificity. This result argues against the model that PFs exclusively function independently and explains the context-dependency of PF activities observed in multiple studies.

Chapter 3

Discussion and future directions

Work presented in this chapter includes a part of the manuscript under review "Stoeber, S., Godin, H., Xu, C., & Bai, L. (2024). Pioneer Factors: Nature or Nurture.".

3.1 Discussion

3.1.1 Summary

In higher eukaryotes, sequence-specific transcription factors (TFs) on average only bind to <1% of their binding motifs across the genome. One major effector is thought to be nucleosomes, which inhibit TF binding. However, a group of TFs named pioneer factors (PFs) can overcome the nucleosome barrier. Intuitively, due to their abilities of invading into closed chromosomes in vivo and binding nucleosomal DNA in vitro, PFs should be able to access most of their consensus motifs in the genome. That's indeed the case for PFs in budding yeast, which occupy almost 100% of their strongest motifs. However, in mammalian cells, even PFs can only bind to a small subset of their binding motifs, and their bindings exhibit cell type specificity. These properties are critical for PFs in higher eukaryotes to function as "master regulators" for cell differentiation / reprogramming because they allow PFs to "pioneer" alternative genes in different cells. How PFs achieve such binding specificity is not well understood.

Here, using a novel synthetic biology approach named "ChIP-ISO", we study the binding specificity of a classic PF, FOXA1, in A549 human lung carcinoma cells. We find that within the same sequence background, FOXA1 binding is strongly enhanced by co-factors including AP-1

and CEBPB. In contrast, chromatin background, including heterochromatin marks, play a minor role in FOXA1 binding. We also show that such cooperativity is physiologically important: it at least partially accounts for the differential binding of FOXA1 in a few different cell types.

3.1.2 Significance of this study

Our work provides a few major insights:

First, our findings go against the classic definition of pioneer factors (PFs)

Traditionally, PFs like FOXA1 are thought to bind and open compacted chromatin without the help of any other factors[118]. In fact, whether a transcription factor (TF) can function independently was proposed to be a key criterion to differentiate it between a PF and a non-PF. Although in recent years, some TFs were proposed to affect PF binding[119,188], these studies have been highly controversial and intensely debated[119,189]. For example, some papers propose that steroid receptors (SR) affect FOXA1 binding through assisted loading, while others argue that they bind independently. To that point, a Viewpoint article recently published in Nature Reviews Genetics clearly demonstrates a lack of consensus in the field on the basic principles of PFs, including the role of co-factors on PF binding[190]. Our study provides clear and convincing evidence supporting that FOXA1 heavily relies on and cooperates with co-factors AP-1 and CEBPB to bind, challenging the classic concept of PF. This is especially striking considering that FOXA1 was the first factor that the name “pioneer factor” was coined and has since been widely studied in the context of its pioneering ability. Our study also gives mechanistic insights into the cell-type-specific function of PFs, which has not been extensively covered in existing literature.

Second, our finding reveals an unexpected relation between heterochromatin and PF binding

H3K9me3-marked heterochromatin was shown to impede the binding of PF Oct4, Sox2 and Klf4[80]. This conclusion was generalized to all PFs and widely accepted, though other PFs have not been extensively tested. In particular, although anti-correlations between PF binding and heterochromatin have been observed, no studies to our knowledge have manipulated the epigenomic landscape surrounding single PF motifs to measure the effect on PF binding. We addressed this critical gap in the literature by employing ChIP-ISO, CRISPRi, and neural network analyses, leading to the conclusion that heterochromatin does not directly impede FOXA1 binding. Our study reveals that this long-recognized concept of heterochromatin impediment to PF binding needs to be revisited on a case-by-case basis. We are confident that this novel finding will spark interest and ignite discussion about the influence of chromatin landscape on PF binding in the field.

Third, a slew of innovative methods is used in this study

We designed and implemented the novel ChIP-ISO method, which allows us not only to dissect the causal relation between PF binding and various genetic features, but also to evaluate the contribution from each feature quantitatively. Compared to commonly used perturbation methods (e.g., gene knock-out, inhibition of epigenetic writers), the ChIP-ISO method avoids global change in chromatin and gene expression and allows us to probe the TF binding rules in essentially wild-type cells at physiological TF levels. Several other state-of-the-art approaches were also developed and used in this work, including EMSA-seq, CRISPRi-mediated epigenetic editing, neural network, etc. Importantly, the combination of these in vivo, in vitro, and in silico methods provides a new experimental framework (**Figure 3-1**) for dissecting the binding rules of

many chromatin factors, and we expect others in the field will be interested in utilizing and adapting our methods for their own studies in the future.

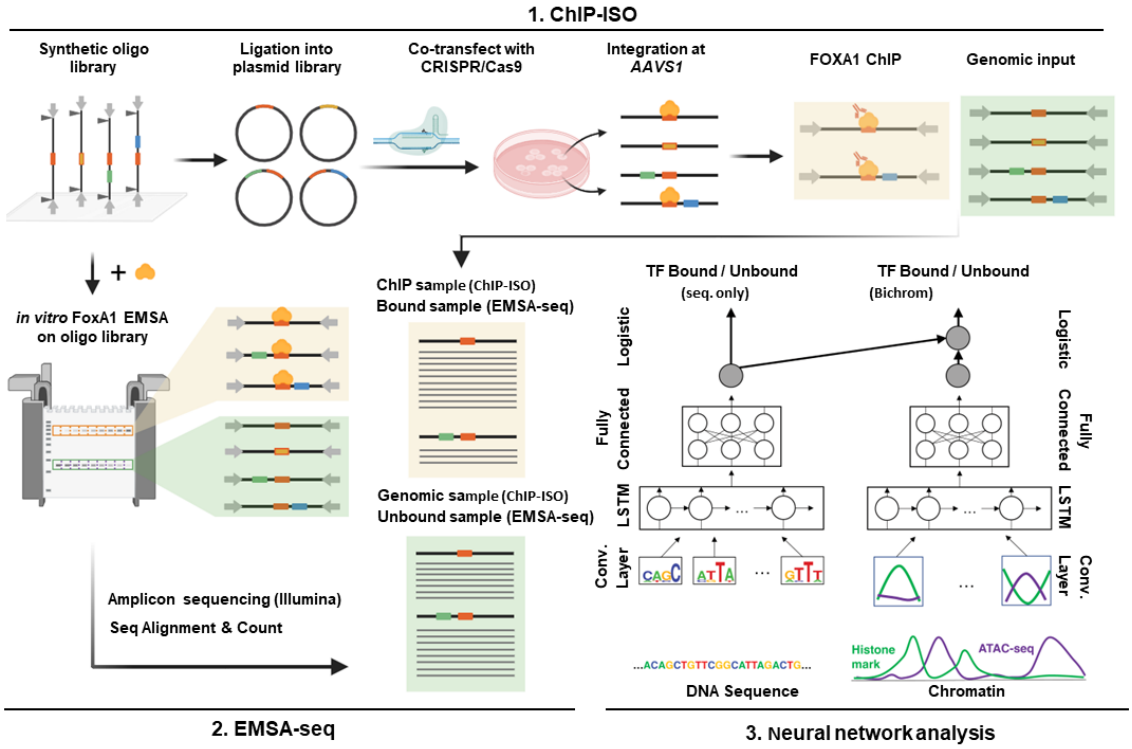


Figure 3- 1: Summary of innovative methods used in this study.

3.1.3 Discussion: PF binding and co-factors

Although PFs can act independently in the most upstream of gene-regulatory networks based on their canonical definition, recent studies have revealed an emerging role that co-factors play in modulating PF binding. The archetypical PF FOXA1 is capable of binding and opening compacted chromatin alone in vitro[118]. Consistent with this property, early studies showed that FOXA1 initiates chromatin opening in vivo, which is required for recruiting steroid receptors (SRs) in breast and prostate cancer cells[83,115,191,192]. Contrary to these results, researchers have

reported that SRs can also recruit FOXA1 to a subset of sites in the genome[119], though controversy still persists[188,189]. In addition, several other TFs including CEBPA[193], GATA4[114] and PDX1[120] have also been found to affect the genome-wide distribution of FOXA1/2. Our systematic study characterized AP-1 and CEBPB as co-factors that strongly enhance FOXA1 binding, and revealed the binding cooperativity between FOXA1 and AP-1. These studies demonstrate that despite being a PF, a proportion of FOXA1's binding events are assisted by a wide range of co-factors. Besides FOXA1, dependency on co-factors to bind was also observed for other PFs. Pluripotency factors OCT4, SOX2 and KLF4 are PFs that participate in the reprogramming of fibroblasts to induced pluripotent stem cells (iPSCs)[80], during which a large fraction of each factor's genomic binding events are dependent on the availability of the other two factors[194]. In other cell types, OCT4 and SOX2's genomic occupancy can be directed by PARP-1[195], OTX2[196] and TFAP2A[176]. Another putative PF PU.1[197] shows extensive cooperative binding with various lineage-determining TFs during myeloid and lymphoid development[27,198-200]. Therefore, dependency on co-factors for binding to a subset of loci may be a commonly existing phenomenon for PFs.

PF binding events assisted by co-factors are usually cell-type specific. PFs FOXA1/2 play crucial roles in endoderm development and were thought to broadly bind and prime lung-, pancreas-, and liver-specific enhancers in foregut endoderm before lineage specification starts[201,202]. However, when FOXA1/2 binding is mapped throughout the course of differentiation, a new set of binding sites was gained after pancreas specification. FOXA1/2 binding to these gained sites is dependent on a pancreas-specific TF PDX1[120]. Furthermore, this type of PF binding events likely carries important lineage-specific functions during development. The majority of FOXA1/2-bound pancreas-specific enhancers require PDX1 for recruiting FOXA1/2 and are unprimed before pancreas induction[120]. Similarly, in the hematopoietic system, cooperative binding between PU.1 and macrophage- or B-cell-specific TFs establishes cell-type

specific PU.1 binding, which drives cell-type specific gene expression and determines the two different cell fates[27]. Pluripotency factors OCT4 and SOX2 also promote the formation of multipotent neural crest, in which their genomic targets are modified by a tissue-specific TF TFAP2A to create a neural crest epigenetic landscape[176]. Additionally, consistent with the observation that clusters of low affinity binding motifs are often found inside developmental enhancers[174,175], these co-factor-dependent PF binding sites tend to contain suboptimal PF motifs based on our and other studies[120,195]. Altogether, these findings suggest a model where PFs can bind to sites with optimal motifs by themselves and prime chromatin, whereas co-factors are required to direct them to a subset of sites containing weak motifs to mediate cell-type and lineage-specific functions.

3.1.4 Discussion: PF binding and epigenetic landscape

Though PFs are able to bind to motifs within closed DNaseI-resistant chromatin, the influence of histone modifications on PF binding remains unclear. PFs can bind to closed chromatin lacking histone modifications, which suggests that PFs do not require open histone modifications, such as those enriched at active promoters and enhancers, for binding[80,114]. There is some evidence of PF binding enrichment on nucleosomes modified by H3K4me1 and H3K4me2, but it remains unclear if this modification is absolutely required for PF binding to these sites[203-205].

In contrast, some studies have found that PFs are unable to access binding sites located in heterochromatin regions enriched in H3K9 methylation. For example, one study found that the reprogramming factors OCT4, SOX2, KLF4, and c-MYC exhibit reduced binding to certain sites in human fibroblasts (Fib) compared to human embryonic (ES) or induced pluripotent stem (iPS) cells, which was correlated with H3K9me3 modification of these sites in Fib cells[80]. Indeed,

knocking down H3K9 methyltransferases in Fib cells led to increased binding of the reprogramming factors to these sites[80]. Other studies have recapitulated the finding that knocking down or inhibiting H3K9 methyltransferases or overexpressing H3K9 demethylases can improve the reprogramming efficiency of differentiated cells into pluripotent or totipotent cells, which could be due to improved binding of reprogramming factors to their demethylated target sites[206-210].

The binding of other PFs, including FOXA1 and FOXA2, has also been observed to be anti-correlated with H3K9me2 and H3K9me3 modifications[114,203]. However, a recent study revealed that FOXA1 is able to access its binding sites within heterochromatin and cause chromatin decompaction of these regions through its intrinsic ability to phase separate into FOXA1 condensates[211]. This behavior could explain why FOXA1 binding appears to be mutually exclusive with H3K9 methylation throughout the genome. We have also found that FOXA1 binding seems to be unaffected by the surrounding landscape. For example, upon ectopically writing H3K9me3 to a FOXA1-bound region using CRISPRi, FOXA1 binding remained unchanged. Additionally, we found that moving a 200 bp region containing an unbound FOXA1 motif from its endogenous heterochromatin locus to a euchromatic locus does not result in increased FOXA1 binding. Together, these results suggest that heterochromatic regions may be permissive to the binding of certain PFs, like FOXA1. However, more studies will need to be performed to clarify if and how histone modifications influence the binding of individual PFs.

3.2 Future directions

3.2.1 Optimization and modification of ChIP-ISO

While the ChIP-ISO method has proven successful in disentangling factors affecting FOXA1 binding specificity in this dissertation, several aspects of the method, such as integration efficiency, required experimental material and reproducibility, could potentially benefit from improvement. Furthermore, the method can be modified to investigate other scientific questions related to chromatin and gene regulation.

Using the landing pad strategy and site-specific recombinases for genome integration

Currently, CRISPR-Cas9 is used to integrate the sequence library into the AAVS1 site. This method is subject to the following limitations: 1) The off-target activity of the Cas9 nuclease cannot be ignored[212]; 2) The integration efficiency, which is a major limiting factor for the size of the synthetic library, could potentially be enhanced; 3) The library sequence could be integrated into more than one alleles of the AAVS1 locus in a single cell, which may impede the study of reporter gene expression; 4) The current strategy cannot be easily applied to integrating the library into other genomic loci, especially heterochromatin regions (discussed below).

The implementation of the landing pad strategy and the utilization of site-specific recombinases could provide a solution to these limitations. Based on the work of Zhang et al. (2023)[213], the synthetic sequence library for ChIP-ISO can be integrated with the following steps (**Figure 3-2**). First, a landing pad containing a blasticidin resistance gene (*BlaR*) driven by a EF1α promoter, with a PhiC31 integrase attP site in between, is integrated into the genomic locus of interest using CRISPR-Cas9 followed by homology-directed repair (HDR). An isogenic cell line with one copy of the landing pad integrated at the right locus is generated as the chassis cell

line. Second, the library plasmids are constructed, consisting of the synthetic sequence library, an optional reporter gene (e.g. EGFP) driven by a core promoter, as well as a PhiC31 integrase attB site followed by a promoterless puromycin resistance gene (*PuroR*). Third, the library plasmids are co-transfected with a plasmid expressing PhiC31 integrase into the chassis cells. Through the integrase-mediated recombination, the plasmid library is integrated into the locus of interest at the attP site. Consequently, the cells switch from blasticidin resistant to puromycin resistant.

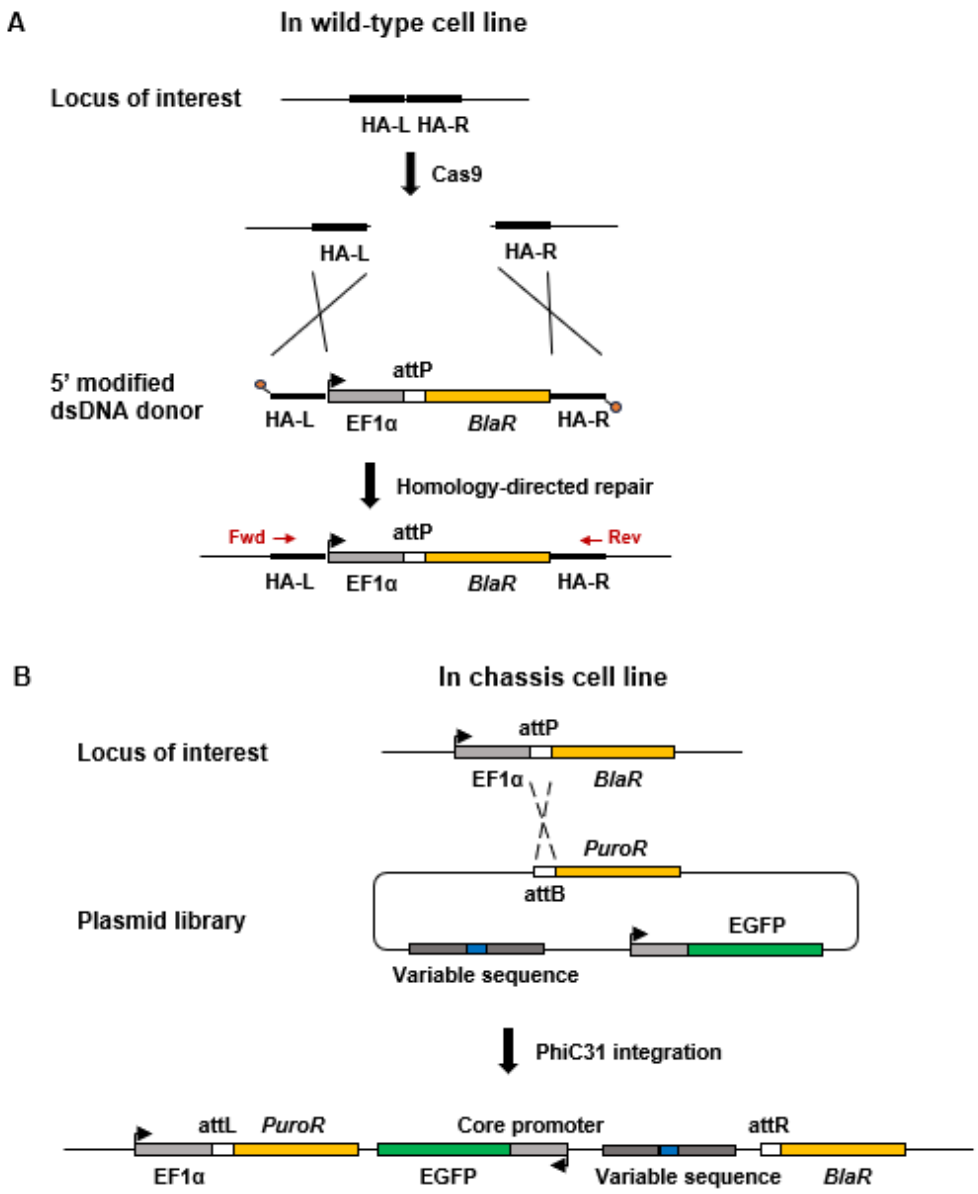
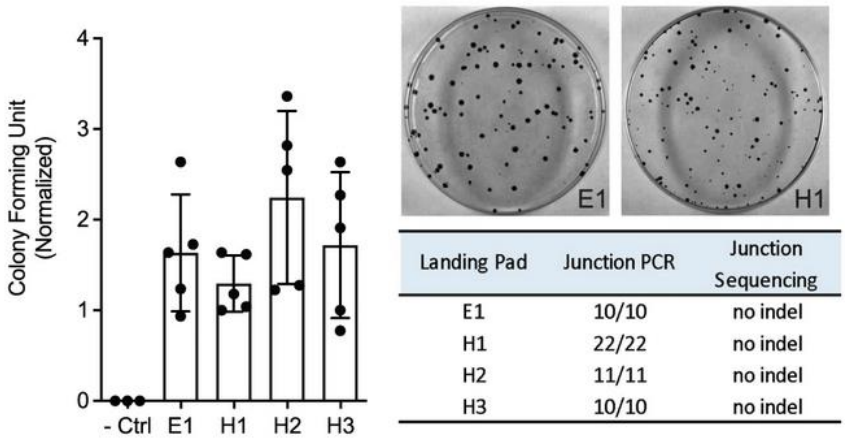


Figure 3- 2: Workflow for using the landing pad strategy and site-specific recombinases for genome integration.

Though off-target activities have also been reported for site-specific recombinases[214], usage of short recombination sites instead of long homologous arms allows the high-fidelity selection design. The *PuroR* will only be expressed when the plasmid library is correctly integrated into the landing pad. Indeed, Zhang et al. reported an on-target integration rate of

100% in their work[213] in human HCT116 cells (versus ~85% in this dissertation). They also detected no off-target integration at endogenous pseudo attP sites. More intriguingly, they achieved an integration efficiency of ~0.4%, which is markedly higher than the efficiency of current CRISPR-Cas9 method (~0.05%). The recent discovery of more potent large serine recombinases[215] may further enhance the efficiency. Promisingly, these findings will point to a ChIP-ISO system that demands less labor and allows for higher library complexity.

The advantages of the new integration system may not outweigh the additional need to construct a chassis cell line when only studying TF binding at the AAVS1 site. However, it is more versatile if integration into other genomic loci is desired. The editing efficiency of CRISPR-Cas9 can be highly variable depending on the design of guide RNA sequences and homologous arms, as well as chromatin configurations of the target loci, such as heterochromatin[216]. For each genomic locus of interest, the CRISPR-Cas9 design needs to be fine-tuned, which still cannot guarantee sufficient integration efficiency. In contrast, the PhiC31 integrase mediates integration into different euchromatin and heterochromatin regions with consistently high efficiency (**Figure 3-3**)[213]. In addition, it allows the same plasmid library to be used for different genomic loci containing the same landing pad, while the CRISPR-Cas9 system requires constructing new plasmid library for each distinct locus. These properties demonstrate that the integrase-based landing pad strategy can be more easily adapted to various genomic loci to study the effect of chromatin background on TF binding. Besides its high on-target integration rate, high efficiency and versatility, the new strategy also ensures only one copy of library plasmid is integrated in each cell, enabling the quantification of reporter gene expression using fluorescence activated cell sorting (FACS).



Landing Pad	Junction PCR	Junction Sequencing
E1	10/10	no indel
H1	22/22	no indel
H2	11/11	no indel
H3	10/10	no indel

Figure 3- 3: Consistently high integration efficiency into euchromatin (E1) and heterochromatin (H1, H2 and H3) regions by PhiC31 integrase.

Picture from Ref[213].

The current integration method can be modified in additional ways. Based on the same landing pad strategy, several other designs employing Cre recombinase[217] or flippase[218] can serve as alternatives to the integrase-based system. A less labor-intensive method to integrate the plasmid library into a number of specific genomic loci has also been reported[219]. In brief, barcoded landing pads are randomly inserted into different genomic locations using lentivirus, creating a pool of cell lines in which the landing pad locations are determined by inverse PCR. Then the plasmid library is integrated into these loci with site-specific recombinases. This method, however, does not allow for precise targeting of the genomic loci of interest.

Cleavage under targets and tagmentation with integrated synthetic oligonucleotides (CUT&Tag-ISO)

With ChIP-ISO, ChIP is utilized to measure TF binding to the integrated DNA sequence library. Since it was developed 40 years ago[220], ChIP has been the most commonly used method for profiling the binding sites of TFs and other chromatin-associated proteins in the genome.

However, it is well-known that ChIP method suffers from a few limitations, including low efficiency, high background, low resolution, demanding in starting materials and sequencing depth, requirement for extensive optimization, biases resulting from formaldehyde cross-linking and sonication, and widely reported "hyperChIPable artifacts[57,173,221-223]". In recently years, emerging technologies such as cleavage under target and release using nuclease (CUT&RUN[56,224]) and Cleavage Under Targets and Tagmentation (CUT&Tag[57]) (**Figure 3-4**) have been reported. They have major advantages over ChIP, including low background, usage with low cell numbers, low sequencing depth requirement, free of cross-linking and chromatin fragmentation, and simple procedures[222,225]. Therefore, they could potentially be applied to improve the current ChIP-ISO method.

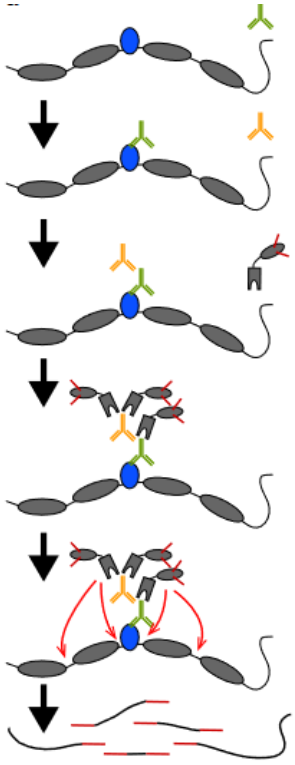


Figure 3- 4: Workflow for CUT&Tag.

Picture from Ref [57].

However, the new protocol may not work well if the ChIP step is simply replaced by CUT&RUN or CUT&Tag. In the ChIP-ISO protocol, ChIP is followed by amplicon sequencing, which involves PCR amplification of a DNA fragment (> 200 bp) containing the intact variable region in the sequence library. This step is compatible with ChIP (average sheared chromatin length > 200 bp[226]), but not CUT&RUN or CUT&Tag (sizes of most TF-binding fragments generated by protein A-MNase or protein A-Tn5 cleavage < 120 bp[56,57]). On the other hand, a different amplicon sequencing strategy that takes advantage of Tn5 tagmentation and usage of barcodes for library sequences could be compatible with CUT&Tag. This strategy also relies on that with CUT&Tag, fragmented DNA remains inside the nucleus rather than gets released into the supernatant in the CUT&RUN protocol[225]. The detailed design and steps of this amplicon sequencing strategy are modified from the work of Liu et al. (2020[136]) with improvement and will be discussed below.

In CUT&Tag-ISO, each sequence in the DNA library will be assigned with a unique barcode. In the plasmid library, the barcodes will be placed 100 bp to 200 bp downstream of the variable library sequences to reduce perturbation to TF binding and occurrences of Tn5 tagmentation downstream of the barcodes. The barcodes will be followed by 4 bp of random sequences (named as unique molecular identifier, UMI) and a constant primer-binding sequence. The UMIs are intended for distinguishing tagmentation events happening at the same position in different copies of the same library sequence (see below), and each library sequence will be associated with one unique barcode and multiple UMIs. To construct the plasmid library (**Figure 3-5**), first, each barcode will be synthesized together with its corresponding sequence in the synthetic oligonucleotide library. Second, the synthetic oligonucleotide library will be PCR amplified, during which UMIs and the constant primer-binding sequence will be added to the 5' end of the reverse primer, so that they can be incorporated downstream of the barcodes. Third, the resulting PCR product will be cloned into a plasmid backbone. Lastly, the obtained plasmids will

be cut between the variable library sequences and the barcodes, and a 100 to 200-bp constant background flanking sequence will be inserted.

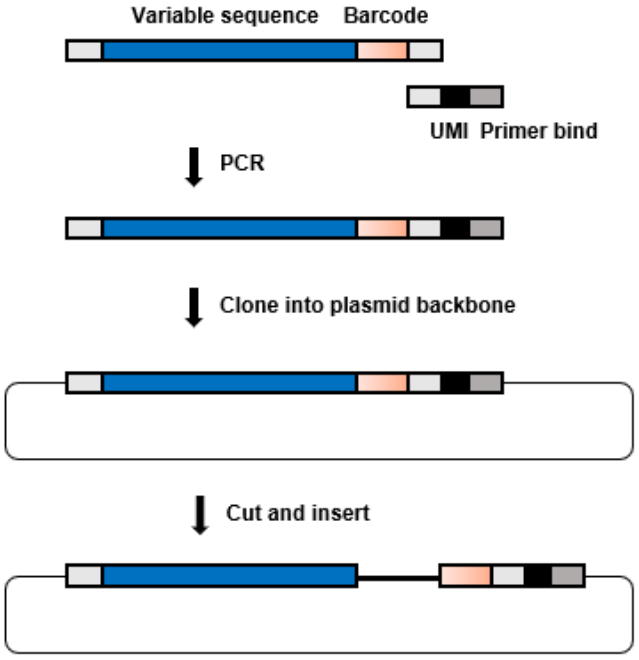


Figure 3- 5: Construction of plasmid library for CUT&Tag-ISO.

The sequences in the library plasmids will be integrated into a fixed locus in the genome, followed by CUT&Tag assay, in which the Tn5 transposase cleaves DNA near the binding sites of the TF of interest and adds two different adaptors to the 5' ends of the two DNA strands in the cleavage sites (tagmentation[225,227]) (**Figure 3-6**). Next, tagmented genomic DNA will be purified, and subject to amplicon PCR. A pair of primers that anneal to the added adaptors and the constant primer-binding sequence downstream of UMI respectively will be used to amplify the region between the primer-binding sequence and the nearest upstream tagmentation event. Because both adaptors could be added to the 5' end downstream of the cleavage sites by tagmentation[228], two pairs of primers are needed to amplify the regions separately in the two situations. The products from the two PCR reactions will be pooled, and used for next-generation paired-end sequencing.

The sequencing reads from two ends of each molecule will recover different information about the molecule, including tagmentation position (from "tagmentation read"), barcode of the library sequence and UMI (from "barcode read") (**Figure 3-7**). To quantify TF binding to a specific library sequence, all the independent tagmentation events corresponding to its barcode will be compiled, and multiple tagmentations at the same position will be distinguished by UMIs to exclude PCR duplicates. The total number of independent tagmentation events will be normalized by the abundance of that library sequence in the cell library as well as the sum of tagmentations in the entire experiment, generating a score for TF binding strength.

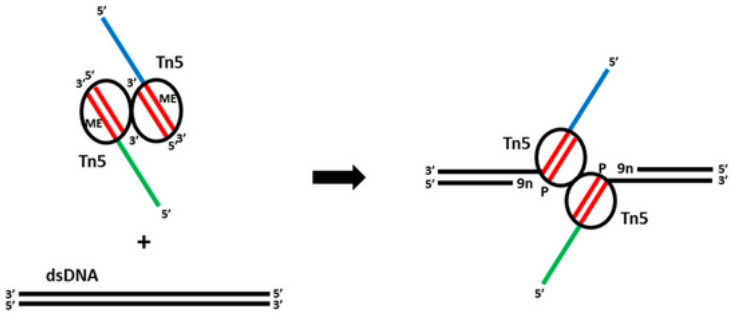


Figure 3- 6: The mechanism of Tn5 tagmentation of DNA.

Picture from Ref [227].

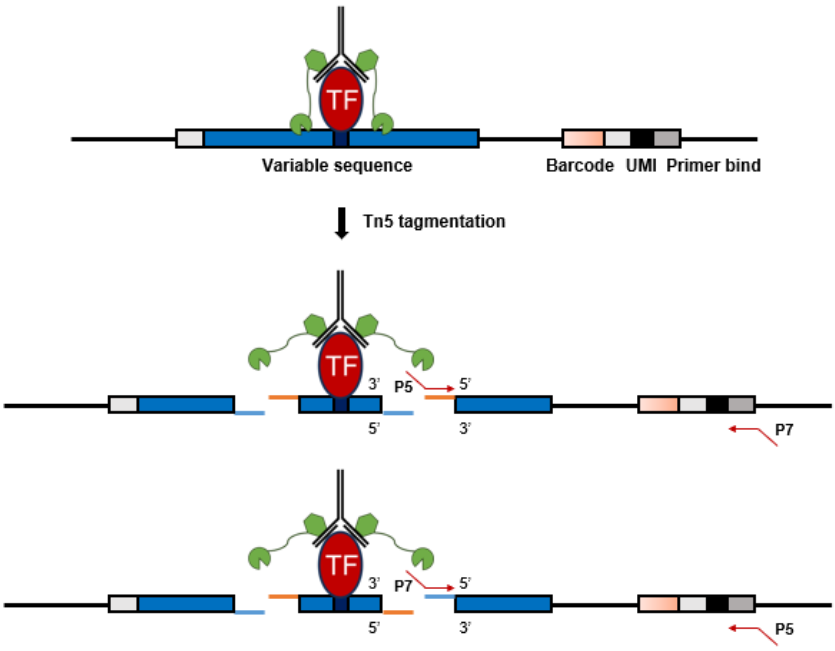


Figure 3- 7: Workflow for CUT&Tag-ISO.

Besides the benefits from the CUT&Tag method, CUT&Tag-ISO have two additional advantages over ChIP-ISO. First, only 2x50bp sequencing is required (ChIP-ISO uses 2x150bp), which reduces the cost and make it easier for pooling with other sequencing libraries. Second, the distribution of tagmentation events on the library sequences (**Figure 3-8**) may contain information about different TF binding modes. Actually, fragment sizes from CUT&RUN experiments have been used to distinguish between direct PF-DNA contact and PF binding to nucleosomes[229,230]. It is possible that such information is also embedded in CUT&Tag-ISO data.

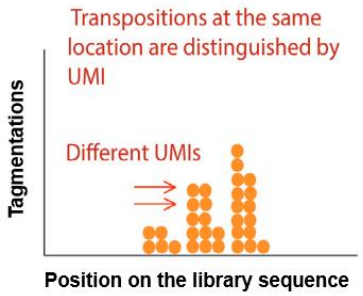


Figure 3- 8: An imaginary distribution of tagmentation events on a library sequence.

Adapted from Ref[136].

The assay for transposase-asseccsible chromatin with integrated synthetic oligonucleotides (ATAC-ISO).

The system for genome integration of synthetic library in ChIP-ISO also provides an attractive platform for highly parallel measurement of chromatin state such as nucleosome occupancy, histone modifications and chromatin accessibility on synthetic sequences. While ChIP-ISO can be directly applied to measure histone modifications, and a similar system using micrococcal nuclease (MNase) to measure nucleosome occupancy has been reported[93], application to assessing chromatin accessibility has not been explored. The assay for transposase-accessible chromatin using sequencing (ATAC-seq) is an increasingly popular method for profiling chromatin accessibility[231,232]. Because the same Tn5 tagmentation process is involved in ATAC-seq and CUT&Tag, the design in CUT&Tag-ISO could be directly applied to ATAC-ISO (**Figure 3-9**).

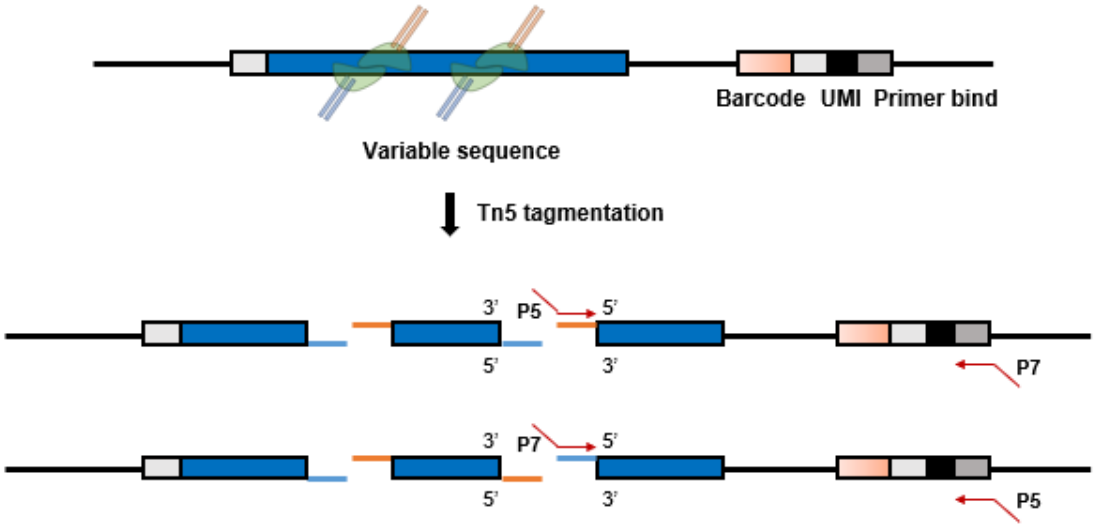


Figure 3- 9: Workflow of ATAC-ISO.

Moreover, analyzing the distribution of tagmentation events from ATAC-ISO data could potentially reveal more information than CUT&Tag-ISO. Additional details that could be learnt regarding the chromatin state include: 1) length of the chromatin accessible region; 2) presence of nucleosomes; 3) footprints from TF binding[233] (**Figure 3-10**). Therefore, ATAC-ISO could potentially be a powerful tool for comprehensive profiling of chromatin state on integrated synthetic library, which is worth further study.

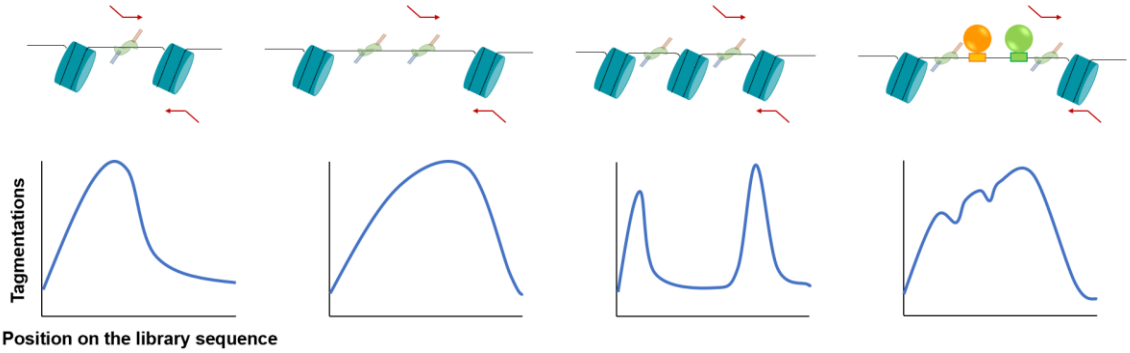


Figure 3- 10: Analyzing the distribution of tagmentation events from ATAC-ISO data could potentially reveal additional chromatin information.

3.2.2 Characterizing the molecular mechanism of FOXA1 and AP-1's binding cooperativity.

This dissertation establishes that FOXA1 and AP-1's binding is cooperative both *in vivo* and *in vitro*, though the detailed molecular mechanism remains unilluminated. As discussed earlier in this chapter, their binding synergy could be caused by DNA-dependent protein-protein interaction and/or nucleosome-mediated cooperativity. We plan to test and characterize these two potential mechanisms using a range of biochemical, genetic and structural biology methods.

Testing if FOXA1 and AP-1 interact and if the interaction is enhanced by DNA.

To directly test the protein-protein interaction between FOXA1 and AP-1 *in vivo*, co-immunoprecipitation (co-IP) will be performed. A new cell line (A549-CX5) in which acute degradation of endogenous FOXA1 protein can be induced by $dTAG^V$-1 molecule[234] will be generated for co-IP and other experiments planned afterwards. The A549 cell line with Dox-inducible A-FOS expression (A549 ePB tet-on A-FOS, reported in this dissertation) will be used as the background cell line. Each allele of the endogenous *FOXA1* gene will be fused with a $FKBP12^{F36V}$ tag and a V5 tag through CRISPR-Cas9 mediated knock-in. In absence of $dTAG^V$-1 (no FOXA1 degradation), tagged FOXA1 will be immunoprecipitated using an antibody against the V5 tag, and all AP-1 subunits that potentially co-precipitate with FOXA1 will be analyzed separately by western blot[235]. As negative controls, co-IP will be conducted with IgG, as well as in +$dTAG^V$-1 condition (FOXA1 is depleted). This experiment will reveal if FOXA1 and AP-1 interact with each other and if FOXA1 binds specific AP-1 subunits.

If FOXA1 and AP-1 interact, to investigate if their interaction is enhanced by DNA binding, FOXA1 mutants whose DNA binding activities are disrupted will be tested. Two FOXA1 mutants will be used: the NH-mut (N216AH220A) with its DNA sequence-specific binding disrupted, and the RR-mut (R262R265A) with its nonspecific DNA binding disrupted[236]. These two mutants, along with the wild-type FOXA1, will be fused with an HA tag, and expressed individually under the control of the cumate-inducible system[237] in the A549-CX5 cell line (**Figure 3-11**). In + $dTAG^V$-1 and + cumate condition, the endogenous FOXA1 protein will rapidly depleted and the HA-tagged wild-type or mutant FOXA1 will be expressed exogenously. To test the interaction between AP-1 and exogenous FOXA1, co-IP will be performed using an antibody against the HA tag, and AP-1 subunits will be analyzed by western blot. Co-IP in +$dTAG^V$-1 and - cumate condition will be used as the negative control. This experiment will

reveal how the disruption of FOXA1's DNA binding activity alters its interaction with AP-1 *in vivo*, which indicates the dependency of their interaction on DNA binding.

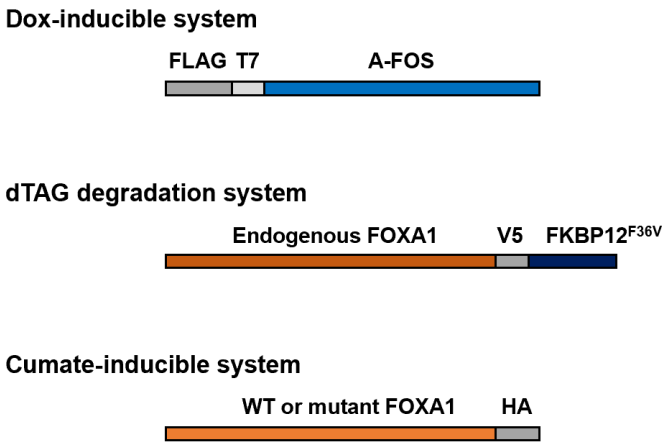


Figure 3- 11: Summary of genetic systems applied to A549-CX5 cell line and expression of HA-tagged WT or mutant FOXA1.

Investigating the basis of FOXA1 and AP-1's interaction

If the experiments above confirm that the cooperativity between FOXA1 and AP-1 relies on their DNA-enhanced protein-protein interaction, we plan to further characterize their interaction domain *in vitro* and then confirm it *in vivo*. To better determine the configurations of FOXA1 and AP-1 motif pair used in the DNA template for *in vitro* assays, a preliminary screen will be conducted to identify the motif spacings and orientations that allow for cooperative binding between FOXA1 and AP-1 *in vitro* using the method consecutive affinity-purification systematic evolution of ligands by exponential enrichment (CAP-SELEX[66]). The DNA library used in CAP-SELEX can be composed of 40-bp random sequences as described in the original protocol[66], or native genomic sequences (e.g. *CCND1e*) with FOXA1 and AP-1 motifs positioned in different orientations and spacings. Motif configurations enriched by CAP-SELEX will be

confirmed by EMSA. Notably, cooperative TF binding identified by both methods could result from direct protein-protein interaction, or indirect interaction through DNA conformational changes[238]. If FOXA1 and AP-1 cooperativity *in vitro* follows similar rules as *in vivo*, i.e. it can happen in both orientations and requires motif proximity, but not a specific spacing, a few motif spacings with each orientation will be selected for the DNA template in the following experiments. Otherwise, specific motif configurations will be selected.

The interaction domain between FOXA1 and AP-1 will then be characterized using two strategies. First, cross-linking mass spectrometry (CLMS[239]) will be implemented to identify proximal amino acid residues when FOXA1 and AP-1 interact *in vitro*, which will indicate the potential interaction domain. Purified FOXA1 and AP-1 proteins will be incubated with individual DNA templates, and subject to CLMS. Alternatively, Cross-linking immunoprecipitation-MS (xIP-MS[240]), which coupled CLMS with co-IP, could be applied to characterize native FOXA1 and AP-1 interaction using whole cell lysates. Second, the structure of FOXA1 and AP-1 co-bound to the DNA templates will be solved by X-ray diffraction or cryogenic electron microscopy (cryoEM). The structure information will directly reveal the basis of FOXA1 and AP-1's interaction and cooperativity, including their contact interface. Besides providing information on the interaction domain of FOXA1 and AP-1, these two experiments may also reveal how they cooperate with different motif configurations, i.e. whether protein-protein interaction is always involved and whether the interaction is specific or promiscuous.

The characterized interaction domain will be further studied *in vivo*. To confirm its role in interaction, amino acid residues in the FOXA1 protein that mediate the contact with AP-1 will be mutated, and the resulting FOXA1 mutants (FOXA1-Δintr) will be fused with an HA tag, and expressed exogenously in the A549-CX5 cell line under the control of the cumate-inducible system. Co-IP will be performed in + dTAG^{V}-1 and + cumate condition using an HA antibody as

described above. The results will be compared to wild-type FOXA1 to test if the protein-protein interaction between FOXA1 and AP-1 is disrupted.

To study how the disruption of FOXA1 and AP-1's interaction influences FOXA1 binding in the genome, ChIP-seq using an antibody against the HA tag will be performed in + dTAG^{V}-1 and + cumate condition for both exogenously expressed wild-type FOXA1 and FOXA1-Δintr. Notably, the mutations that disrupt FOXA1's ability to interact with AP-1 might also change the intrinsic chromatin binding activity of FOXA1, altering FOXA1 occupancy globally. To specifically test how AP-1-dependent FOXA1 binding is affected, FOXA1 binding will be analyzed separately in AP-1-facilitated FOXA1 binding sites (lost peaks) and independent FOXA1 binding sites (unchanged peaks) identified in this dissertation. For each site, the ratio of FOXA1-Δintr binding to wild-type FOXA1 binding will be calculated. Significantly lower ratios in the AP-1-facilitated sites than the independent sites will suggest that the protein-protein interaction between FOXA1 and AP-1 indeed contributes to their binding cooperativity *in vivo* (**Figure 3-12**).

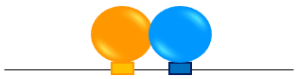


Figure 3- 12: TF binding cooperativity mediated by DNA-enhanced protein-protein interaction.

Test if FOXA1 and AP-1 also cooperate in interaction-independent manners.

Aside from direct protein-protein interaction, cooperative TF binding could also be mediated by interaction-independent mechanisms, including DNA conformational change-mediated and nucleosome-mediated cooperativity[65,185,238] (**Figure 3-13**). Even if the experiments

above may demonstrate that direct contact between FOXA1 and AP-1 contributes to their cooperative binding, it does not rule out the possibility that these two modes of interaction-independent cooperativity also have an effect. To test that, we will utilize the cell line constructed above, in which HA-tagged FOXA1-Δintr is expressed exogenously in the A549-CX5 cell line under the control of the cumate-inducible system. In + dTAG^{V}-1 and + cumate condition, AP-1 and FOXA1-Δintr, which replaces the endogenous FOXA1 protein, co-bind in an interaction-free manner. AP-1 binding will then be inhibited by A-FOS overexpression (+ Dox). FOXA1-Δintr binding genome-wide will be measured by ChIP-seq using an HA antibody in + Dox versus – Dox conditions. Differential binding analysis will be performed to examine if a significant fraction of FOXA1-Δintr binding sites still lose binding in response to AP-1 inhibition. AP-1 co-binding and presence of AP-1 motifs will be confirmed at these lost sites. This experiment will reveal whether and the extent to which interaction-independent mechanisms contribute to the cooperative binding between FOXA1 and AP-1.

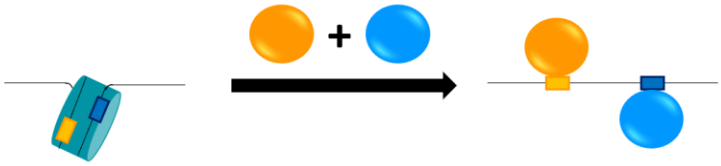


Figure 3- 13: Nucleosome-mediated TF binding cooperativity.

The interaction-independent cooperativity between FOXA1 and AP-1 could be mediated by DNA conformational change and/or nucleosome depletion, whose individual contributions are hard to be decoupled *in vivo*. Whether alteration in DNA structure could facilitate the cooperativity can be more easily tested *in vitro*. The solved structure of FOXA1 and AP-1 co-bound to the DNA templates may provide insights into such mechanisms. EMSA can also be done using purified FOXA1-Δintr and AP-1 protein, along with naked DNA template, to test if cooperative effects still exist in the absence of direct contact between the two proteins.

If DNA conformational changes indeed contribute to the interaction-independent cooperative binding of FOXA1 and AP-1, it will be difficult to rigorously evaluate the existence and effects of nucleosome-mediated cooperativity *in vivo*. Several experiments and analysis may provide some indications for this issue. First, whether FOXA1-Δintr and AP-1 show more cooperativity on nucleosomal DNA than naked DNA can be studied using EMSA. Second, the configurations of FOXA1 and AP-1 motif pairs in the lost FOXA1-Δintr binding sites upon AP-1 inhibition can be analyzed. In line with nucleosome-mediated cooperativity, the two motifs should be located within 150 bp of distance while motif proximity and specific orientation are not strictly required[75,241]. Third, the nucleosome occupancy in the lost FOXA1-Δintr binding sites before and after AP-1 inhibition can be measured by micrococcal nuclease (MNase) assay to test if nucleosome occupancy is high without the co-binding of FOXA1-Δintr and AP-1 and gets reduced with that.

3.2.3 Investigating the function of FOXA1 and AP-1's binding cooperativity in chromatin opening and enhancer selection.

While this dissertation has focused on studying how the binding specificity of FOXA1 is impacted by its cooperativity with AP-1, the functional role of this type of binding events is less understood. Given the fact that FOXA1 is an archetypical PF capable of opening closed chromatin[118,158] and AP-1 has also been recently proposed to be a PF[181], we propose that the cooperative binding of these two PFs may enable them to engage with and open certain closed chromatin regions that cannot be accessed and remodeled by a single PF. We define this type of pioneering activity, which relies on the collaboration of two PFs, as co-pioneering activity (**Figure 3-14**). Based on our results that cell-type specific binding of FOXA1 in A549 cells is partially caused by high AP-1 expression, we further propose that opening of chromatin by the

co-pioneering activity of FOXA1 and AP-1 is context-specific and may lead to the priming and activation of cell-type specific enhancers.

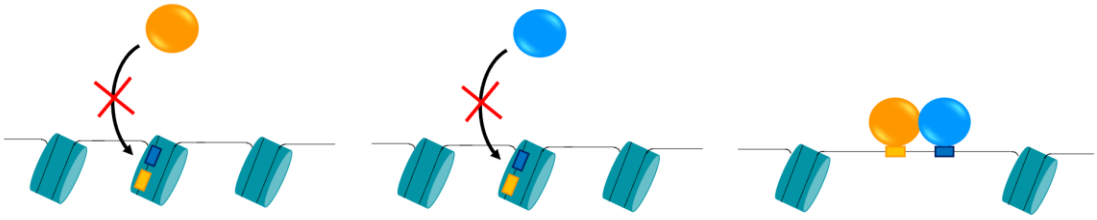


Figure 3- 14: Model of a co-pioneering event initiated by the cooperative binding of FOXA1 and AP-1.

To test this hypothesis rigorously, a cell line with no endogenous FOXA1 and AP-1 expression is required as the background cell line. FOXA1 and AP-1 will be expressed exogenously by two different inducible expression systems, so that their expressions can be induced separately or together. Examination of FOXA1 and AP-1's binding and activities in this system will be most compatible with the classic definition of PFs[76]. However, due to the ubiquitous expression of AP-1 and differential presence of its seven subunits, it might be hard to find such background cell lines. Alternatively, we could use A549 cells in which both FOXA1 and AP-1 are expressed, and study how loss of their functions affects the existing chromatin state. Applying this strategy, we will take advantage of the A549-CX5 cell line constructed in the previous session. In this cell line, rapid degradation of FOXA1 and inhibition of AP-1 binding can be induced by adding dTAG^{V}-1 and Dox respectively.

We will first seek to comprehensively characterize cooperative binding events of FOXA1 and AP-1 in A549 cells. Using the A549-CX5 cell line, FOXA1 ChIP-seq will be performed in +/- Dox and - dTAG^{V}-1 conditions, and FOSL2 ChIP-seq will be performed in - Dox and +/- dTAG^{V}-1 conditions. Differential binding analysis will be conducted for both experiments to identify AP-1-facilitated FOXA1 binding sites and FOXA1-facilitated AP-1 binding sites. The

common loci between these two sets of sites will serve as cooperative binding loci of FOXA1 and AP-1.

To test if co-pioneering by FOXA1 and AP-1 is responsible for chromosome opening in a substantial fraction of their cooperative binding loci, chromatin state in these loci will be assessed in four different conditions (+/- Dox and +/- dTAG^{V}-1) in the A549-CX5 cell line. Genome-wide chromatin openness will be measured by three different methods: ATAC-seq, MNase-seq and DNase-seq. A locus co-pioneered by FOXA1 and AP-1 will have high chromatin openness (high ATAC signal, low MNase signal and with DNase sensitivity) only in the - Dox and - dTAG^{V}-1 condition, while having low chromatin openness in the other three conditions. Loci with other combinations of chromatin openness may also exist and can be compared to the co-pioneered loci for further evaluation.

Next, we will try to test if co-pioneering by FOXA1 and AP-1 in these loci is partially responsible for cell-type specific chromatin opening and enhancer activation in A549 cells. The chromatin and enhancer landscape in A549-CX5 cells (- Dox and - dTAG^{V}-1) will be compared with two other cell lines separately: HepG2 cells with high expression of FOXA1 and low expression of AP-1 subunits, and K562 cells with relatively high expression AP-1 subunits and low expression of FOXA1. Based on the model that PFs remodel chromatin in a sequential process[154], we will conduct ATAC-seq as well as H3K4me1 and H3K27ac ChIP-seq, which identify accessible regions, primed enhancers and active enhancers[242] respectively, in these three cell types. Differential analysis will be performed in A549-CX5 versus HepG2 and A549-CX5 versus K562 cells to characterize A549-enriched, shared and HepG2/K562-enriched accessible regions and enhancers. For each category, the percentage of regions that overlap with FOXA1 and AP-1's co-pioneering loci will be calculated. Significantly higher percentages observed for A549-enriched accessible regions and enhancers will indicate that the co-pioneering activity of

FOXA1 and AP-1 mainly contributes to A549-specific chromatin opening events and enhancer selection.

Appendix A

Methods

Cell Lines. Wild-type A549 human lung carcinoma epithelial cells, a gift from Dr. Yanming Wang, were maintained in Ham's F-12K (Kaighn's) Medium (Gibco 21127022) supplemented with 10% FBS (Gibco 16000044) and 1% Penicillin-Streptomycin (Gibco 15070063). Wild-type HepG2 human liver cancer cells and MCF-7 human breast cancer cells were obtained from ATCC, and maintained in Dulbecco's Modified Eagle Medium (DMEM) (Gibco 10569044) supplemented with 10% FBS (Gibco 16000044) and 1% Penicillin-Streptomycin (Gibco 15070063). All cells were cultured at 37°C in a humidified incubator with 5% CO_2. Cells at passage number one were thawed and passaged at least an additional two times prior to experimental usage.

Synthetic Oligonucleotide Design. The ChIP-ISO sequence library comprises 3,203 different sequences, each 229bp in length with 193bp variable regions and 18bp primer-binding regions on the two sides, each containing a BsaI or BbsI recognition site (different subsets of oligos use different primers and cutting sites) (**Figure 2-4**). The synthetic oligo library was ordered from Agilent (Product #G7220A).

To design the ISO library with *CCND1e* variants, a 193bp region from *CCND1e* (chr11:69,654,913-69,655,105) was selected (**Figure 2-12**), and an internal BbsI cutting site was mutated to distinguish it from the native *CCND1e*. The library sequences were designed using a MATLAB program developed previously in the lab[243]. Each FOXA1 motif was mutated, reversed, or converted into consensus. For the motifs of the other eight co-factors, the 1 to 3 most consensus bases were mutated to their complementary bases to maintain GC content, while

avoiding interfering with neighboring motifs (checked by MEME "FIMO"[244], Version 5.5.4). Some combinations of FOXA1 and co-factor motif variations were also included.

To design the ISO library containing native sequences with FOXA1 and TF cobinding in **Figure 2-25**, overlaps between the top 30,000 genomic FOXA1 peaks from our FOXA1 ChIP-seq data and TF ChIP-seq peaks from ENCODE were identified using BEDTools "intersect intervals"[245] (Version 2.30.0). The number of overlapping regions was divided by 30,000 (number of total FOXA1 peaks) to give the percent of FOXA1 peaks overlapping the TF. The enrichment of the TF motif within the 30,000 FOXA1 peaks was calculated using MEME "AME"[246] (Version 5.5.4). A selection of TFs having a large percentage of FOXA1 peak overlaps and high motif enrichment in FOXA1 peaks were chosen for further analysis. FOXA1 and TF motifs within FOXA1 / TF overlapped regions were identified using FIMO with Position Weight Matrices (PWM) obtained from JASPAR or CIS-BP. These regions were filtered to select only for sequences containing a FOXA1 and TF motif within 8-30bp measured from the center of each motif. 10 to 20 regions were included in the synthetic library for each TF. Additionally, for each region, versions containing a mutated FOXA1 or TF motif were also included.

To design the ISO library "set 1" in **Figure 2-52**, which contains strong FOXA1 motifs but shows very weak or no FOXA1 binding, we first used FIMO to locate all FOXA1 motifs in the genome and calculated their motif scores based on PWM. A subset of motifs with score >16 and have no overlap with FOXA1 ChIP-seq peaks (evaluated by BEDTools "intersect intervals") were selected for the library. For set2 where weak FOXA1 motifs are associated with strong ChIP-seq peaks, we first sorted FOXA1 ChIP-seq peaks based on their intensities using deepTools2[247]. Among the top 50% of the peaks, the corresponding genomic sequences (peak center +/- 100bp) were retrieved using bedtools getfasta (Version 2.30.0), and FOXA1 motifs

within these sequences were identified by FIMO. Sequences that contain a single motif with the score between 10-14.5 were selected for the library. To design the ISO library containing FOXA1 motifs covered by different epigenetic marks, FIMO was first used to identify FOXA1 motifs within FOXA1 ChIP-seq peaks and ChIP-seq peaks of different histone modifications. FOXA1 motifs present in both FOXA1 and H3K9me3/H3K27me3 peaks represent FOXA1-bound motifs within these repressive regions. FOXA1 motifs associated with FOXA1 peaks but absent in all histone modification datasets were labeled as FOXA1-bound motifs within unmarked regions. FOXA1-unbound motifs within repressive or unmarked regions were labeled similarly. A random subset of 50-60 sequences within each of these categories were included in the ChIP-ISO library, with the exception of FOXA1-bound motifs within H3K9me3-marked regions, in which all 21 regions were included. Each 193bp sequence was designed to include the genomic region surrounding the centered FOXA1 motif.

Plasmid Construction. The plasmid library backbone was derived from pAAVS1-Nst-MCS, which was a gift from Knut Woltjen (Addgene plasmid # 80487 ; http://n2t.net/addgene:80487 ; RRID:Addgene_80487). A ~2kb human genomic region containing the *CCND1e* (chr11:69,653,809-69,655,876) was cloned between the PacI and SalI cutting sites. The two endogenous BsaI cutting sites on the resulting plasmid were mutated. The 193bp *CCND1e* sequence (chr11:69,654,913-69,655,105) was replaced by two BsaI cutting sites. The resulting plasmid named pCX1.10 was used as the backbone plasmid for ChIP-ISO plasmid library construction.

The plasmid expressing Cas9 and gRNA was derived from eSpCas9(1.1), a gift from Feng Zhang (Addgene plasmid # 71814 ; http://n2t.net/addgene:71814 ; RRID:Addgene_71814).

sgRNA-T2 (5'-GGGGCCACTAGGGACAGGAT-3') was cloned between the two BbsI cutting sites. The resulting plasmid was named pCX3.10.

To construct the plasmid for A-FOS overexpression, A-FOS sequence was PCR amplified from plasmid CMV500 A-FOS, a gift from Charles Vinson (Addgene plasmid # 33353 ; http://n2t.net/addgene:33353 ; RRID:Addgene_33353), and cloned into Xlone-GFP plasmid[248] (a gift from Dr. Lance Lian), replacing the GFP gene between the KpnI and SpeI cutting sites. The resulting plasmid was named pCX4.1.

To design the piggyBac gRNA-containing vector used to randomly integrate *CCND1e*-targeting gRNAs into the genome, we followed a protocol described previously, with the following changes[249]. Two *CCND1e*-targeted gRNAs were designed using CHOPCHOP at a distance of roughly -300bp from the first FOXA1 motif and +300bp from the third FOXA1 motif. Oligonucleotides (IDT) corresponding to these two gRNA sequences were cloned into pGEP179_pX330K (Addgene 137882) and pX330S-2 (Addgene Kit 1000000055) according to the kit instructions. These gRNA-containing vectors were assembled by Gibson assembly into a single entry vector for Gateway cloning into pGEP163 (Addgene 137881), resulting in the plasmid pGEP163_CCND1_U2_D4.

Generation of the Plasmid Library for ChIP-ISO. The synthetic oligonucleotide library was resuspended in TE buffer, pH 8.0, and diluted with water to 10 nM for PCR amplification. For each 1,000 types of oligonucleotides, 32 µl of 10 nM diluted synthetic library was amplified in a 400 µl PCR reaction (final template concentration 800 pM) for 13 cycles, using NEBNext Ultra II Q5 master mix (NEB M0544S). The PCR product was purified using Amicon Ultra-2mL 50K centrifugal filter (Millipore UFC205024) to exchange the PCR solution

for 1X NEB CutSmart buffer. DNA concentration was estimated by agarose electrophoresis. ~1.8 μg of the amplified library was digested with 60 U of BsaI/BbsI at 37°C overnight, followed by adding 30 U of extra BsaI/BbsI and digestion at 37°C for another two hours. The digestion products were purified with AMPure XP beads (Beckman Coulter MSPP-A63880) with a beads to DNA ratio of 1.8. 5 μg of pCX1.10 plasmid was digested with 120 U of BsaI at 37°C overnight, followed by adding 30 U of extra BsaI and digestion at 37°C for another two hours. 10 U of CIP was added into the digestion reaction, followed by incubation at 37°C for one hour to dephosphorylate 5'-ends. The linearized plasmid backbone was purified with E.Z.N.A. cycle pure kit (Omega D6492-02). 600 ng linearized plasmid backbone and 48 ng of digested library (molar ratio of 1:3) were ligated with 5 μL (2000 U) of T4 DNA ligase (NEB M0202S) in a 100 μL reaction. The ligation reaction was incubated at 16°C overnight, purified with E.Z.N.A. cycle pure kit and eluted with 30 μL of water. The purified ligation product was transformed into 5-alpha electrocompetent *E. coli* (NEB C2989) via electroporation. In each electroporation reaction, 25 μL of electrocompetent cells was transformed with 2.5 μL of purified ligation product. Adequate number of electroporation reactions were done to produce at least ~100,000 colonies per 1,000 types of oligonucleotides. The *E. coli* cells were then pooled and grown overnight with ampicillin selection, followed by plasmid extraction using the E.Z.N.A. plasmid DNA maxi kit (Omega D6922-02).

Generation of the Cell Library for ChIP-ISO. 9.06×10^6 wild-type A549 cells were plated per 15 cm dish in 22.5 mL of Ham's F-12K (Kaighn's) Medium supplemented with 10% FBS and 1% Penicillin-Streptomycin. 24 hours later, cells in each dish were transfected with 7.875 μg of pCX3.10 plasmid (expressing gRNA and Cas9), 23.625 μg of library plasmids, 96.75 μL of lipofectamine 3000 reagent (ThermoFisher L3000015) and 63 μL of P3000 reagent, which were diluted in Opti-MEM. 8-10 hours post-transfection, the media was replaced with fresh

Ham's F-12K (Kaighn's) Medium supplemented with 10% FBS and 1% Penicillin-Streptomycin. 48 hours post-transfection, the cells in each dish were dissociated from the dish and splitted into two 15 cm dishes with Ham's F-12K (Kaighn's) Medium supplemented with 15% FBS, 1% Penicillin-Streptomycin and 600 μg/mL of G418. Media changes were performed every three to four days while the G418 selection was kept. When cell colonies were visible, the number of colonies was estimated by counting colonies inside randomly sampled grids on the dish under the microscope. Adequate number of transfection reactions were done, which produced ~92,000 colonies for 3,203 types of sequences. The cells were then dissociated from the dishes, disaggregated, pooled and plated in new 15 cm dishes with Ham's F-12K (Kaighn's) Medium supplemented with 10% FBS, 1% Penicillin-Streptomycin and 300 μg/mL of G418. The pooled cell library was maintained and expanded for ChIP.

Chromatin Immunoprecipitation (ChIP). ChIP was performed with the cell library for ChP-ISO following a standard ChIP protocol. To fix protein-DNA interactions, formaldehyde (Ricca Chemical Company RSOF0010250A) was added to 2 X 10^7 adherent log-phase cells to a final concentration of 1% and incubated for 10 minutes at room temperature. For FOXA1 and histone modification ChIP-ISO samples, 1.6 X 10^8 cells and 8 X 10^7 cells were fixed, respectively, in individual plates of 2 X 10^7 cells for each replicate. Cross-linking was quenched by addition of glycine to a final concentration of 0.125 M and incubated for 5 minutes at room temperature. Cells were washed twice with cold 1X DPBS. Cells were scraped into cold 1X DPBS and pelleted at 4 °C. In some cases, cell pellets were snap frozen using liquid nitrogen and stored at -80°C until ready to proceed. Fresh or thawed cell pellets were lysed by incubating cells in 2.5 mL cell lysis buffer (5 mM PIPES pH 8.0, 85 mM KCl, 0.5% NP-40) supplemented with protease inhibitor cocktail (Sigma-Aldrich P8340) for 10 minutes on ice. Cell nuclei were pelleted at 4°C and lysed in 150 μL nuclei lysis buffer (50 mM Tris-Cl pH 8.0, 10 mM EDTA,

1% SDS) supplemented with protease inhibitor cocktail for 10 minutes on ice. Chromatin was fragmented in Diagenode Pico with a circulating water bath at 4°C using the Shear and Go Easy Mode setting for 3 cycles (30 seconds on followed by 30 seconds off). Sonicated chromatin was centrifuged to remove cell debris and residual SDS precipitate. The supernatant containing sheared chromatin was pooled across all replicates, a 50 µL input DNA sample was reserved, and the remaining pool was split again into eight (FOXA1) or four (histone modification) chromatin samples. In some cases, supernatant containing sheared chromatin was snap frozen using liquid nitrogen and stored at -80°C until ready to proceed.

For each ChIP sample, 20 µL Magna ChIP™ Protein A+G Magnetic Beads (Sigma-Aldrich 16-663) was washed four times with 1X DPBS supplemented with 5 mg/mL BSA and subsequently crosslinked to 5 µg antibody (Anti-FOXA1 antibody: GeneTex, GTX100308; Anti-Fra2 antibody: Cell Signaling Technology, 19967S; Anti-H3K9me3 antibody: abcam, ab8898; and Anti-H3K27me3 antibody: abcam, ab6002) for two hours at 4°C. Antibody-crosslinked magnetic beads were washed four additional times with the DPBS/BSA solution. Each chromatin sample except the input was incubated with the washed antibody-crosslinked magnetic beads for two hours at 4°C. Next, the magnetic beads were washed five times with LiCl wash buffer (100 mM Tris pH 7.5, 500 mM LiCl, 1% NP-40, 1% sodium deoxycholate) and once with 1X TE buffer (10 mM Tris-HCl pH 7.5, 0.1 mM Na_2EDTA) at room temperature. Immunoprecipitated chromatin was eluted from the magnetic beads by incubating in IP elution buffer (1% SDS, 0.1 M $NaHCO_3$) for 1 hour at 65°C. The collected supernatant (containing immunoprecipitated chromatin) and the reserved input sample were both incubated with 40 µg RNase A and NaCl to a final concentration of 0.37 M overnight at 65°C. The next day, 80 µg proteinase K was added to the ChIP sample and 400 µg was added to the input sample and both were incubated at 55°C for 2 hours. The ChIP and input DNA was purified via phenol-chloroform extraction. At the DNA

elution step, the individual ChIP DNA pellets were resuspended together to achieve a more concentrated sample.

Prior to any downstream applications, the success of the ChIP reaction was determined via qPCR using various positive, negative, and locus-specific primer pairs. qPCRs were performed using the Agilent AriaMx Real-Time PCR System with the SYBR Green optical module (Emission 516.0 nm, Excitation 462.0 nm). For ChIP-seq, the same ChIP protocol was used, but with the following modifications. For each ChIP biological replicate, only 2 X 10^7 adherent log-phase cells were fixed. 50 µL of sheared chromatin was reserved per replicate as the genomic input sample. At the DNA elution step, each ChIP pellet was resuspended separately.

Amplicon Sequencing for ChIP-ISO. To amplify the integrated ChIP-ISO library sequences while excluding other genomic DNA, including native *CCND1e*, we performed two rounds of PCR amplification with a BbsI digestion step in between (**Figure 2-9**). The primer pairs for the first round of PCR contain regions annealing to the *CCND1e* (outside the 193bp library sequence) at the 3'-ends, partial Illumina TruSeq adaptor sequences at the 5'-end and 0-3 random nucleotide spacers in between to increase sequence complexity. Primers with different numbers of spacers are mixed in equimolar ratio for the first round of PCR. Preliminary PCR tests were performed to decide the optimal cycle number that keeps the PCR reactions in exponential phase. We used 23-25 cycles for our first round of PCR. For the first round of PCR, 30 µl of ChIP DNA was amplified in a 100 µl PCR reaction using NEBNext Ultra II Q5 master mix. The PCR products were purified with AMPure XP beads with a beads to DNA ratio of 0.9. 15 µL of the purified PCR product was digested with 20 U of BbsI in a 30 µL reaction at 37°C for two hours. The digestion products were purified with AMPure XP beads with a beads to DNA ratio of 0.9, followed by the second round of PCR. The primer pairs for the second round of PCR contain the

rest of the Illumina TruSeq adaptor sequences and sample indexes. For the second round of PCR, 2 µl of purified digestion product was amplified in a 50 µl PCR reaction for 8 cycles using NEBNext Ultra II Q5 master mix. The PCR products were purified with AMPure XP beads with a beads to DNA ratio of 0.8. Quality control was conducted with TapeStation (Agilent). 30 million paired-end 150bp reads were obtained for each ChIP-ISO and input sample using a NextSeq 2000 (Illumina) instrument. Demultiplexing was performed using DRAGEN BCL Convert (v3.8.4).

Sequencing Data Analysis for ChIP-ISO. Raw sequencing reads were filtered using fastp with default settings[250] (Version 0.23.2), and the first three nucleotides were trimmed from each read using cutadapt[251] (Version 4.4) to remove the 0-3 random nucleotide spacers introduced by the amplicon primers. Processed forward and reverse reads were merged into single reads based on their overlapping regions using NGmerge[252] (Version 0.1). Merged reads were aligned to a FASTA file containing all ChIP-ISO library sequences (including their reversely ligated versions) using BWA-MEM2[253] (Version 2.2.1). BAM alignments containing at least 2 mismatched nucleotides were filtered out using BAMtools[254] (Version 2.4.0). The number of filtered reads aligning to each ChIP-ISO library sequence was counted for each ChIP and input sample and normalized to the total number of sequencing reads for each sample. ChIP-ISO signal was calculated by dividing the normalized number of ChIP counts by the normalized number of input counts. Any sequence having fewer than 1000 input counts was excluded from further analyses, resulting in FOXA1 ChIP-ISO signals for 1,882 sequences. ChIP-ISO signal is reported and plotted as the average of two independent biological ChIP-ISO replicates.

Low-throughput ChIP-qPCR test of binding on single integrated sequences. The overall process was the same as ChIP-ISO. Instead of the synthetic oligonucleotide library, single

synthetic sequences were cloned into the pCX1.10 plasmid backbone. Wild-type A549 cells were transfected with the resulting plasmids individually, together with pCX3.10 plasmid. For each synthetic sequence, the cell colonies were pooled together after G418 selection, and expanded for ChIP. To measure TF binding to the integrated synthetic sequences and the native *CCND1* enhancer separately, quantitative PCR (qPCR) was conducted with primer pairs that can distinguish between the integrated and the native sequences. Locked nucleic acids (LNAs) were incorporated into the primers to increase specificity.

ChIP-seq and Data Analysis. Sequencing library was constructed by NEBNext ultra II DNA library prep kit (NEB E7103L). 50 million paired-end 50bp reads were obtained for each ChIP and input sample using a NextSeq 2000 instrument. Paired-end reads were filtered using fastp with default settings and subsequently aligned to the human reference genome (hg38) using BWA-MEM2. Resulting BAM files were filtered for MAPQ scores > 20 using SAMtools[255] (Version 1.8). Mapped regions within the ENCODE Blacklist were excluded from further analysis[256] (hg38 Version 2). Read coverage was obtained separately for input and ChIP samples using deepTools "bamCoverage", with a bin size of 10bp[247] (Version 3.5.1), and visualized in IGV (Version 2.8.12) or the UCSC Genome Browser. ChIP peaks were called from pooled ChIP replicates using MACS2 callpeak with the default settings[257] (Version 2.1.1.20160309). Heatmaps and intensity profiles were generated using computeMatrix, plotHeatmap, and plotProfile functions in deepTools2[247] (Version 3.5.4).

RNA-seq. Cells were lysed by Trizol (ThermoFisher 15596026), extracted by 0.2 volume of chloroform, followed by adding equal volume of 100% ethanol. RNeasy kit (Qiagen 74104) was then used to purify the RNA. 3 μg of purified RNA was treated with 2 U of RNase-free DNase I, and purified again with RNeasy kit. Sequencing library was constructed by the Illumina

Stranded mRNA Prep kit (Illumina 20040532). 30 million paired-end 50bp reads were obtained for each RNA-seq sample using a NextSeq 2000 instrument. Data analysis was conducted based on a protocol from Batut et al. 2021[258].

A-FOS-Related Experiments and Data Analysis. To construct A549 ePB tet-on A-FOS cell line, $6x10^5$ wild-type A549 cells were plated in a 6-well plate well in Ham's F-12K (Kaighn's) Medium supplemented with 10% FBS and 1% Penicillin-Streptomycin. 24 hours later, cells were transfected with 0.72 μg of piggyBac transposase plasmid (System Biosciences PB210PA-1) and 1.78 μg of pCX4.1 plasmid using Lipofectamine 3000 transfection reagent. 12 hours post-transfection, the media was replaced with fresh medium. 48 hours post-transfection, the cells in the well were dissociated from the dish and splitted into two 6-well plate wells with Ham's F-12K (Kaighn's) Medium supplemented with 15% FBS, 1% Penicillin-Streptomycin and 10 μg/mL of blasticidin. Media changes were performed every three to four days while the blasticidin selection was kept. When cell colonies were visible, the cells were then dissociated from the wells, disaggregated, pooled and maintained in Ham's F-12K (Kaighn's) Medium supplemented with 10% FBS, 1% Penicillin-Streptomycin and 5 μg/mL of blasticidin. The expanded cells were subject to immunofluorescence, ChIP-seq and RNA-seq.

To identify lost, unchanged and gained FOXA1 ChIP-seq peaks in +Dox versus -Dox conditions, differential binding analysis was conducted with Bioconductor "DiffBind"[259] (Version 3.18) using default settings. The regions within each category were converted from BED to FASTA format using BEDTools "GetFastaBed", and FOXA1 and AP-1 motif scanning was performed inside these regions using MEME "FIMO". To identify proximal genes of FOXA1 ChIP-seq peaks, MEME "T-Gene"[260] (Version 5.5.4) was first used to predict target genes for each category of peaks. The predicted target genes, whose distances to the corresponding ChIP-

seq peaks are smaller than 100 kb, are selected as the proximal genes for further analysis. Gene ontology (GO) analysis on the proximal genes of the lost peaks was performed by "Metascape"[261] (Version 3.5.20230501) using default settings.

Recombinant Protein Expression and Purification. Mouse FOXA1 (UniProtKB: P35582) fused to an N-terminal 6x-histidine tag was expressed in BL21(DE3)pLysS *E. coli* cells (Novagen 69388-3) at 37°C for 3 hours using the bacterial expression plasmid pET-28b-FOXA1, a gift from K. Zaret, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA[89]. Harvested bacterial cells were resuspended and lysed by sonication in P300 buffer (50 mM sodium phosphate pH 7.0, 300 mM NaCl, 5 mM 2-mercaptoethanol, 1 mM benzamidine). Following extraction of soluble proteins, insoluble material containing FOXA1 was resuspended and sonicated in P300 buffer with 7 M urea added. Solubilized FOXA1 was isolated using Ni-NTA chromatography (GoldBio H-350-25), and further purified by Source S cation-exchange chromatography (Cytiva 17-0944-01). FOXA1 protein was stored in 8 mM HEPES pH 7.5, 80 mM NaCl, 8 mM 2-mercaptoethanol, 0.7 M urea, 20% glycerol.

Genes encoding human c-Fos (UniprotKB: P01100) fused to an N-terminal 6x-histidine tag and untagged human c-Jun (UniprotKB: P05412) were subcloned into pST39[262] and pST50Tr[263], respectively, from pST39-F:cJun/6xHis:cFos, a gift from C.M. Chiang, UT Southwestern Medical Center, Dallas, TX[264]. Expression of each protein was carried out separately in Rosetta2(DE3)pLysS *E. coli* cells (Novagen 70951) at 37°C. Soluble proteins were extracted as described for FOXA1, and insoluble materials containing c-Fos and c-Jun were processed separately. c-Fos was solubilized by resuspension and sonication in T100 buffer (20 mM Tris-HCl pH 7.5, 100 mM NaCl, 5 mM DTT). Following centrifugation, clarified extract was dialyzed into P300 buffer containing 7 M urea, and c-Fos was partially purified from it using

Ni-NTA chromatography. Insoluble material containing c-Jun was washed three times in T100 buffer, followed by solubilization in 20 mM Tris-HCl pH 7.5, 1 mM EDTA, 1 mM DTT, 6 M guanidine-HCl. Refolding of c-Fos/c-Jun heterodimers was performed by stepwise dialysis as described by Ferguson and Goodrich, 2001[265] and purified by cobalt metal-affinity chromatography (Talon resin, Clontech 635652). cFos/cJun was stored in 18 mM Tris-HCl pH 7.5, 90 mM NaCl, 9 mM 2-mercaptoethanol, 20% glycerol. Proteins were analyzed by SDS-PAGE.

Electrophoretic Mobility Shift Assay with Sequencing (EMSA-seq) A pooled equimolar mixture of the ChIP-ISO synthetic oligonucleotide library was PCR-amplified and purified via agarose gel purification (Thermo Scientific K0691). The EMSA protocol was adapted from Garcia et al. 2019[78]. Briefly, 100 nM ChIP-ISO synthetic oligonucleotide library was incubated with 50X non-specific competitor DNA and 0-60 nM recombinant mouse 6xHis-FOXA1 in binding buffer (10 mM Tris-HCl pH 7.5, 1 mM $MgCl_2$, 10 µM $ZnCl_2$, 50 mM KCl, 3 mg/mL BSA, 10% glycerol, and 1 mM DTT) at room temperature for 30 minutes. Free and FOXA1-bound library sequences were separated on a 7.5% non-denaturing polyacrylamide gel (Bio-Rad 4561026) run in 1X Tris-Glycine at 200V at room temperature for 30 minutes. Gels were stained with 1 µg/mL Ethidium Bromide (Invitrogen 15585011) in 1X Tris-Glycine for 10 minutes at room temperature. Stained gels were visualized with a Bio-Rad GelDoc Go Imaging System using the Ethidium Bromide setting (**Figure 2-44**). The FOXA1-bound and -unbound DNA bands were excised at each FOXA1 concentration, and the DNA was eluted from each polyacrylamide gel slice following a User-Developed Protocol for extraction of DNA fragments from polyacrylamide gel using the QIAGEN QIAquick Gel Extraction Kit (QIAGEN 28704). The gel-extracted DNA was PCR-amplified and purified using AMPure XP Beads (Beckman

Coulter A63880) with a 0.9x bead cleanup ratio. The library was created in the same manner as the ChIP-ISO library.

30 million paired-end 150bp reads were obtained for FOXA1-bound and -unbound DNA samples at each protein concentration using a NextSeq 2000 instrument. The sequencing data was processed and analyzed in the same manner as the ChIP-ISO datasets. The number of filtered paired-end reads aligning to each ChIP-ISO library sequence was counted for each bound and unbound sample and normalized to the number of sequencing reads for each sample. The ratio of FOXA1-bound DNA to total input (FOXA1-bound + FOXA1-unbound) DNA was calculated for each synthetic oligonucleotide library sequence at each FOXA1 concentration and normalized to the corresponding FOXA1-bound ratio of negative control CCND1e-FOXA1^{all_mut} (Index: 22). The resulting negative-normalized FOXA1-bound ratios were further normalized to the highest ratio across all FOXA1 concentrations, forcing ratios to fall between 0-1 for ease of analysis. For simplicity, these values are called the "ratio bound *in vitro*." To correct for systematic error, values from the first EMSA-seq replicate were adjusted such that the two replicates approximated $r = 1$. To determine the FOXA1 concentration at which the ratio bound *in vitro* falls into the linear range for most sequences, these values were plotted for all library sequences in the top-down *CCND1e* FOXA1 motif category. For each sequence, the data points were fit with the Hill slope equation $Y=B_{max}*X_h/(K_{dh} + X_h)$, where X = FOXA1 concentration and Y = ratio bound *in vitro*, to model specific FOXA1 binding. For all quantitative analyses, the ratio of each synthetic oligonucleotide library sequence bound by FOXA1 *in vitro* was approximated by its FOXA1-bound ratio at 15 nM FOXA1, which was extrapolated by averaging its FOXA1-bound ratios at 10 nM and 20 nM FOXA1. This FOXA1 concentration was chosen because it falls into the linear range of the FOXA1 binding curves of most sequences.

FOXA1 and AP-1 Co-Binding Electrophoretic Mobility Shift Assays (EMSAs).

Three *CCND1e* DNA templates were designed to include different FOXA1 and AP-1 motif mutants. FOXA1^{12_mut} has mutations in the two upstream FOXA1 motifs, FOXA1^{all_mut} has mutations in all three FOXA1 motifs, and FOXA1^{12_mut}, AP-1^{mut} has mutations in the two upstream FOXA1 motifs and single AP-1 motif. These DNA templates were individually PCR-amplified and purified using a PCR clean-up kit (Omega Bio-tek D6492). In the FOXA1 EMSAs, 100 nM *CCND1e* DNA was incubated with 50X non-specific competitor DNA and 0-200 nM recombinant mouse 6xHis-FOXA1 in DNA-binding buffer (10 mM Tris-HCl pH 7.5, 1 mM $MgCl_2$, 10 µM $ZnCl_2$, 50 mM KCl, 3 mg/mL BSA, 10% glycerol, and 1 mM DTT), with or without 300 nM recombinant human AP-1. Similarly, in the AP-1 EMSAs, up to 400 nM cJun/6His:cFos was titrated into the same buffer/DNA solution, with or without 150 nM FOXA1. The EMSA samples were incubated at room temperature for 30 minutes and separated on a 7.5% non-denaturing polyacrylamide gel (Bio-Rad 4561026) run in 1X Tris-Glycine at 200V at room temperature for 30 minutes. Gels were stained with 1 µg/mL Ethidium Bromide (Invitrogen 15585011) in 1X Tris-Glycine for 10 minutes at room temperature. Stained gels were visualized with a Bio-Rad GelDoc Go Imaging System using the Ethidium Bromide setting (**Figure 2-48A**). For the FOXA1 titration EMSAs, the unbound fraction was calculated at each FOXA1 concentration by normalizing the intensity of the free band to the intensity of the free band at 0 nM FOXA1 ± AP-1 (**Figure 2-48B**). The equivalent calculations were performed for the AP-1 titration EMSAs.

CRISPRi. To integrate the KRAB-dCas9 construct into the AAVS1 locus, 2 X 10^6 cells were co-transfected with 10 µg pT077 (Addgene 137879), 1.5 µg AAVS1 TALEN L (Addgene 59025) and 1.5 µg AAVS1 TALEN R (Addgene 59026) using Lipofectamine 3000 Transfection Reagent (Invitrogen L3000015) according to the manufacturer's protocol. Transfected cells were

transferred to a medium supplemented with 700 μg/mL G418 (Gibco 10131035) 24 hours post-transfection and maintained in this medium to allow for single-cell colony formation (~14 days). Single colonies were picked and seeded into 24-well plates. Colonies were maintained in G418-supplemented medium until they reached sufficient cell density, and those that retained normal cell morphology and growth rate were split for visualization of EGFP and maintenance. To visualize the EGFP expression of each colony, cells were plated into two wells of an 8-well dish (ibidi 80806); one well of cells was induced with 1 μg/mL doxycycline (Sigma-Aldrich D5207) 24 hours after plating and one well was left untreated. 48 hours post-induction, inducible expression of KRAB-dCas9 was confirmed by measuring EGFP expression in induced cells normalized to untreated cells. Colonies with high relative EGFP expression and homogeneity were frozen for storage. The colony with the highest EGFP expression and homogeneity was validated using genotyping PCR to confirm KRAB-dCas9 integration at the AAVS1 locus. To randomly integrate *CCND1e*-targeting gRNAs throughout the genome of KRAB-dCas9 cells, 6 X 10^5 cells were co-transfected with 5 μg of gRNA-containing piggyBac vector and 1 μg of piggyBac transposase plasmid (System Biosciences PB210PA-1) using Lipofectamine 3000 Transfection Reagent. Transfected cells were transferred to a medium supplemented with 700 μg/mL G418 and 10 μg/mL Blasticidin S HCl (Gibco A11139) 24 hours post-transfection and maintained in this medium to allow for single-cell colony formation (~14 days). Constitutive expression of gRNAs was confirmed by measuring the expression of mRFP in a mixed cell population. Anti-Histone H3 (tri-methyl K9) antibody (abcam ab8898) was used to perform H3K9me3 ChIP on the mixed cell population, followed by qPCR to confirm H3K9me3 deposition at the *CCND1e*. Anti-FOXA1 antibody (GeneTex GTX100308) was used to perform FOXA1 ChIP on the mixed cell population, followed by qPCR, to monitor any changes to FOXA1 binding at the *CCND1e* upon H3K9me3 deposition.

Cell Type-Specific FOXA1 and AP-1 Motifs, Binding, and Expression. To identify FOXA1 binding sites that are enriched, shared, or depleted in A549 cells compared to HepG2 / MCF-7 cells, a HepG2 FOXA1 ChIP-seq BED file containing significant peaks from at least two replicates was acquired from the ENCODE Database, sorted by coordinate using BEDTools "sortBED", and pooled with the top 30,000 significant WT A549 FOXA1 ChIP-seq peaks from two replicates we acquired. To determine the number of reads aligning to each region in the pooled BED file across cell types and replicates, the pooled BED file and the corresponding BAM files from two HepG2 and two A549 replicates were input into BEDTools "MultiCovBed" using default parameters. The output from this tool was input into Bioconductor "edgeR"[266] (Version 3.34.0) using default settings to identify regions enriched, shared, or depleted in FOXA1 binding in A549 vs HepG2. Regions with a $\log_2$ fold-change less than -1 were labeled A549-depleted, between -1 and 1 were labeled shared, and greater than 1 were labeled A549-enriched. The regions within each category were converted from BED to FASTA format using BEDTools "GetFastaBed", and each category was input into MEME "FIMO" to identify the number of regions containing FOS or JUN motifs. The percent of regions in each category containing each individual FOS or JUN motif was calculated.

Immunofluorescence. Immunofluorescence experiments were performed according to a protocol from Yoney et al. 2022[267]. The following primary antibodies and dilutions were used: FLAG (mouse monoclonal, Millipore-Sigma, F1804, 1:1000), FOXA1 (rabbit polyclonal, GeneTex, GTX100308, 1:500), and FOSL1 (mouse monoclonal, Santa Cruz Biotechnology, sc-28310, 1:50). The following secondary antibodies and dilutions were used: goat anti-mouse IgG(H+L) (Alexa Fluor 594, ThermoFisher, A-11005, 1:1000), and goat anti-rabbit IgG(H+L) (Alexa Fluor 488, ThermoFisher, A-11008, 1:500).

Immunostained A549 ePB tet-on A-FOS cells were imaged using Leica DMI6000 with Hamamatsu ORCA-R2 C10600 camera and SOLA SE light source. Images were acquired in phase contrast, GFP, and Texas Red channels, and with 40x/1.30 objective. Immunostained wild-type A549, MCF-7 and HepG2 cells were imaged using Zeiss Axio Observer 7 with camera Axiocam 705 mono. Images were acquired in DIC, AF594, AF488 and DAPI channels, and with 20x/0.8 objective. The average fluorescence intensity within the nuclei of each cell in the field was calculated using the Zeiss Bio Apps Gene Expression tool. The measurement area was limited to the cell nucleus, which was detected from the signal in the DAPI-stained channel. Average fluorescence intensity in the green/red was then measured for each cell in the field and normalized to the DAPI intensity in the same cell to correct for differences in cell permeability across cell types.

Neural network architectures: The sequence-only convolutional neural network (CNN) model aims to predict FOXA1 ChIP-seq peaks using DNA sequence input. Briefly, one-hot encoded DNA sequence input of length 240bp is first passed through a 1D convolution layer of 256 filters, where each filter is of size 24 and stride 1. After convolution, the output is processed by ReLU activation and batch normalization. A 1D max-pooling layer of size 15 and stride 15 is then applied to pool the output. The pooled output is fed into a long short-term memory (LSTM) layer to output a 32-length vector. The output vector passes through two dense layers with ReLU activation and Dropout. Finally, a single sigmoid activated linear node outputs the prediction probability.

The Bichrom models aim to assess whether chromatin features positively contribute to predicting FOXA1 binding and use a previously published interpretable bimodal neuron network architecture named Bichrom[171]. Bichrom consists of two independent sub-networks,

corresponding to DNA sequence and chromatin input, respectively. The sequence sub-network is the trained CNN network described above with all the trained weights frozen. The final linear node is replaced by a new linear node activated by a tanh function. The input to the chromatin sub-network consists of the relevant chromatin feature(s) coverage track(s), binned in 20bp bins, across the same 240bp region as DNA sequence input. The chromatin feature input passes through a ReLU activated 1D convolution layer of 15 filters (kernel size 1) and a LSTM layer to output a 5-vector. A tanh activated linear node is then used to get the scalar output. The full Bichrom model works by combining the scalar values from both sub-networks into a sigmoid activated linear node to predict the TF binding label. Three Bichrom models were tested: one trained on DNA-sequence and ATAC-seq features; one trained on DNA-sequence and H3K9me3 features; and one trained on DNA-sequence and ATAC-seq, H3K9me3, H3K27ac, H3K4me1, H3K4me2, and H3K4me3 features. All chromatin features were sourced from ENCODE A549 ATAC/ChIP-seq experiments.

Neural network training: For both model architectures, two chromosomes are held out for validation (chr11) and test (chr17). The sampling strategies differ by the model type. For the sequence-only CNN model, positive sample regions are obtained by randomly shifting 240bp long regions centered by ChIP-seq peak midpoints (-95bp<=shifting distance<95bp). Negative sample regions are sampled from four different sources: 1) flanking negative regions around ChIP-seq peaks (flanking distances: [450, -450, 500, -500, 1250, -1250, 1750, -1750]); 2) accessible regions not overlapping ChIP-seq peaks; 3) non-accessible regions not overlapping ChIP-seq peaks; 4) random regions sampled from the entire genome and not overlapping ChIP-seq peaks. The goal of sampling is to ensure the percentages of accessible regions in both positive and negative samples are the same. Bichrom models use the same positive sample regions as the

sequence-only CNN model, while negative sample regions only consist of random regions sampled from the entire genome and not overlapping ChIP-seq peaks.

Neural network feature attribution: The DeepLift-SHAP implementation from SHAP[166,167] was employed to compute the attribution scores for each trained model. The hypothetical attribution scores were obtained by computing the DeepLift-SHAP score of all possible nucleotide choices at each base pair. Then TF-MoDISco was used to extract globally high-impact sequence patterns with the option -n 50000. The final sequence patterns were then compared to motifs from Cis-BP[268] using Tomtom[269].

DNA shape analysis: The DNA shape scores were computed by DNAShapeR[60,270,271] using default settings. Each type of DNA shape score was plotted around the FOXA1 motif center.

Appendix B

References

1. Mendel, G. Versuche uber Pflanzen-Hybriden. Verh. Naturforsch. Ver. Brunn 4 3-47 English translation in 1901. *JR Hortic. Soc* **26**, 1-32 (1866).
2. Morgan, T.H. Sex Limited Inheritance in Drosophila. *Science* **32**, 120-2 (1910).
3. Avery, O.T., Macleod, C.M. & McCarty, M. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *J Exp Med* **79**, 137-58 (1944).
4. Watson, J.D. & Crick, F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-8 (1953).
5. Crick, F.H. On protein synthesis. in *Symp Soc Exp Biol* Vol. 12 8 (1958).
6. Nirenberg, M.W. & Matthaei, J.H. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad Sci U S A* **47**, 1588-602 (1961).
7. Nirenberg, M. & Leder, P. Rna Codewords and Protein Synthesis. The Effect of Trinucleotides Upon the Binding of Srna to Ribosomes. *Science* **145**, 1399-407 (1964).
8. Khorana, H.G. *et al.* Polynucleotide synthesis and the genetic code. *Cold Spring Harb Symp Quant Biol* **31**, 39-49 (1966).
9. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**, 318-56 (1961).
10. Sharp, P.A. Split genes and RNA splicing. *Cell* **77**, 805-15 (1994).
11. Houseley, J. & Tollervey, D. The many pathways of RNA degradation. *Cell* **136**, 763-76 (2009).
12. Brito Querido, J., Diaz-Lopez, I. & Ramakrishnan, V. The molecular basis of translation initiation and its regulation in eukaryotes. *Nat Rev Mol Cell Biol* (2023).
13. Deribe, Y.L., Pawson, T. & Dikic, I. Post-translational modifications in signal integration. *Nat Struct Mol Biol* **17**, 666-72 (2010).
14. Roeder, R.G. & Rutter, W.J. Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature* **224**, 234-7 (1969).
15. Weinmann, R., Raskas, H.J. & Roeder, R.G. Role of DNA-dependent RNA polymerases II and III in transcription of the adenovirus genome late in productive infection. *Proc Natl Acad Sci U S A* **71**, 3426-39 (1974).
16. Fuda, N.J., Ardehali, M.B. & Lis, J.T. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* **461**, 186-92 (2009).
17. Malik, S. & Roeder, R.G. Regulation of the RNA polymerase II pre-initiation complex by its associated coactivators. *Nat Rev Genet* **24**, 767-782 (2023).
18. Adelman, K. & Lis, J.T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet* **13**, 720-31 (2012).
19. Kwak, H. & Lis, J.T. Control of transcriptional elongation. *Annu Rev Genet* **47**, 483-508 (2013).
20. Porrua, O. & Libri, D. Transcription termination and the control of the transcriptome: why, where and how to stop. *Nat Rev Mol Cell Biol* **16**, 190-202 (2015).
21. Roeder, R.G. The role of general initiation factors in transcription by RNA polymerase II. *Trends in biochemical sciences* **21**, 327-335 (1996).

22. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol* **19**, 621-637 (2018).
23. Chen, X. & Xu, Y. Structural insights into assembly of transcription preinitiation complex. *Curr Opin Struct Biol* **75**, 102404 (2022).
24. Wittkopp, P.J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* **13**, 59-69 (2011).
25. Allen, B.L. & Taatjes, D.J. The Mediator complex: a central integrator of transcription. *Nat Rev Mol Cell Biol* **16**, 155-66 (2015).
26. Lambert, S.A. *et al.* The Human Transcription Factors. *Cell* **172**, 650-665 (2018).
27. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-89 (2010).
28. Gauthier, B.R. *et al.* PDX1 deficiency causes mitochondrial dysfunction and defective insulin secretion through TFAM suppression. *Cell Metab* **10**, 110-8 (2009).
29. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663-76 (2006).
30. Lee, T.I. & Young, R.A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237-51 (2013).
31. Latchman, D.S. Transcription factors: an overview. *Int J Biochem Cell Biol* **29**, 1305-12 (1997).
32. Payvar, F. *et al.* Purified glucocorticoid receptors bind selectively in vitro to a cloned DNA fragment whose transcription is regulated by glucocorticoids in vivo. *Proc Natl Acad Sci U S A* **78**, 6628-32 (1981).
33. Dynan, W.S. & Tjian, R. Isolation of transcription factors that discriminate between different promoters recognized by RNA polymerase II. *Cell* **32**, 669-80 (1983).
34. Parker, C.S. & Topol, J. A Drosophila RNA polymerase II transcription factor binds to the regulatory site of an hsp 70 gene. *Cell* **37**, 273-83 (1984).
35. Sawadogo, M. & Roeder, R.G. Interaction of a gene-specific transcription factor with the adenovirus major late promoter upstream of the TATA box region. *Cell* **43**, 165-75 (1985).
36. Carthew, R.W., Chodosh, L.A. & Sharp, P.A. An RNA polymerase II transcription factor binds to an upstream element in the adenovirus major late promoter. *Cell* **43**, 439-48 (1985).
37. Bram, R.J. & Kornberg, R.D. Specific protein binding to far upstream activating sequences in polymerase II promoters. *Proc Natl Acad Sci U S A* **82**, 43-7 (1985).
38. Brayer, K.J. & Segal, D.J. Keep your fingers off my DNA: protein-protein interactions mediated by C2H2 zinc finger domains. *Cell Biochem Biophys* **50**, 111-31 (2008).
39. Emerson, R.O. & Thomas, J.H. Adaptive evolution in zinc finger transcription factors. *PLoS genetics* **5**, e1000325 (2009).
40. Imbeault, M., Helleboid, P.-Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550-554 (2017).
41. Hayward, A., Cornwallis, C.K. & Jern, P. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. *Proceedings of the National Academy of Sciences* **112**, 464-469 (2015).
42. Heffer, A. & Pick, L. Conservation and variation in Hox genes: how insect models pioneered the evo-devo field. *Annual review of entomology* **58**, 161-179 (2013).
43. Inukai, S., Kock, K.H. & Bulyk, M.L. Transcription factor–DNA binding: beyond binding site motifs. *Current opinion in genetics & development* **43**, 110-119 (2017).

44. Stormo, G.D., Schneider, T.D., Gold, L. & Ehrenfeucht, A. Use of the 'Perceptron'algorithm to distinguish translational initiation sites in E. coli. *Nucleic acids research* **10**, 2997-3011 (1982).
45. Schneider, T.D. & Stephens, R.M. Sequence logos: a new way to display consensus sequences. *Nucleic acids research* **18**, 6097-6100 (1990).
46. Weirauch, M.T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443 (2014).
47. Castro-Mondragon, J.A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic acids research* **50**, D165-D173 (2022).
48. Bailey, T.L., Johnson, J., Grant, C.E. & Noble, W.S. The MEME suite. *Nucleic acids research* **43**, W39-W49 (2015).
49. Jolma, A. *et al.* Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome research* **20**, 861-873 (2010).
50. Berger, M.F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology* **24**, 1429-1435 (2006).
51. Meng, X., Brodsky, M.H. & Wolfe, S.A. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nature biotechnology* **23**, 988-994 (2005).
52. Hellman, L.M. & Fried, M.G. Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions. *Nature protocols* **2**, 1849-1861 (2007).
53. Furey, T.S. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics* **13**, 840-852 (2012).
54. Rhee, H.S. & Pugh, B.F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408-1419 (2011).
55. van Steensel, B., Delrow, J. & Henikoff, S. Chromatin profiling using targeted DNA adenine methyltransferase. *Nature genetics* **27**, 304-308 (2001).
56. Skene, P.J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6**(2017).
57. Kaya-Okur, H.S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun* **10**, 1930 (2019).
58. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research* **22**, 1798-1812 (2012).
59. Arvey, A., Agius, P., Noble, W.S. & Leslie, C. Sequence and chromatin determinants of cell-type–specific transcription factor binding. *Genome research* **22**, 1723-1734 (2012).
60. Li, J. *et al.* Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. *Nucleic Acids Res* **45**, 12877-12887 (2017).
61. Slattery, M. *et al.* Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* **147**, 1270-1282 (2011).
62. Abe, N. *et al.* Deconvolving the recognition of DNA shape from sequence. *Cell* **161**, 307-318 (2015).
63. Reiter, F., Wienerroither, S. & Stark, A. Combinatorial function of transcription factors and cofactors. *Current opinion in genetics & development* **43**, 73-81 (2017).
64. Panne, D. The enhanceosome. *Current opinion in structural biology* **18**, 236-242 (2008).
65. Morgunova, E. & Taipale, J. Structural perspective of cooperative transcription factor binding. *Curr Opin Struct Biol* **47**, 1-8 (2017).
66. Jolma, A. *et al.* DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384-8 (2015).

67. Kim, S. *et al.* DNA-guided transcription factor cooperativity shapes face and limb mesenchyme. *Cell* (2023).
68. Domcke, S. *et al.* Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528**, 575-9 (2015).
69. Barozzi, I. *et al.* Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. *Mol Cell* **54**, 844-857 (2014).
70. Sinha, K.K., Bilokapic, S., Du, Y., Malik, D. & Halic, M. Histone modifications regulate pioneer transcription factor cooperativity. *Nature* **619**, 378-384 (2023).
71. Kim, S. & Shendure, J. Mechanisms of Interplay between Transcription Factors and the 3D Genome. *Mol Cell* **76**, 306-319 (2019).
72. Garcia, D.A. *et al.* An intrinsically disordered region-mediated confinement state contributes to the dynamics and function of transcription factors. *Mol Cell* **81**, 1484-1498 e6 (2021).
73. Kornberg, R.D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98**, 285-294 (1999).
74. Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F. & Richmond, T.J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251-260 (1997).
75. Adams, C.C. & Workman, J.L. Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. *Mol Cell Biol* **15**, 1405-21 (1995).
76. Zaret, K.S. & Carroll, J.S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev* **25**, 2227-41 (2011).
77. Iwafuchi-Doi, M. & Zaret, K.S. Pioneer transcription factors in cell reprogramming. *Genes & development* **28**, 2679-2692 (2014).
78. Magnani, L., Eeckhoute, J. & Lupien, M. Pioneer factors: directing transcriptional regulators within the chromatin environment. *Trends in Genetics* **27**, 465-474 (2011).
79. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861-72 (2007).
80. Soufi, A., Donahue, G. & Zaret, K.S. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell* **151**, 994-1004 (2012).
81. You, J.S. *et al.* OCT4 establishes and maintains nucleosome-depleted regions that provide additional layers of epigenetic regulation of its target genes. *Proc Natl Acad Sci U S A* **108**, 14497-502 (2011).
82. Jozwik, K.M. & Carroll, J.S. Pioneer factors in hormone-dependent cancers. *Nature Reviews Cancer* **12**, 381-385 (2012).
83. Laganiere, J. *et al.* From the Cover: Location analysis of estrogen receptor alpha target promoters reveals that FOXA1 defines a domain of the estrogen response. *Proc Natl Acad Sci U S A* **102**, 11651-6 (2005).
84. Magnani, L., Ballantyne, E.B., Zhang, X. & Lupien, M. PBX1 genomic pioneer function drives ERalpha signaling underlying progression in breast cancer. *PLoS Genet* **7**, e1002368 (2011).
85. Gee, J.M. *et al.* Overexpression of TFAP2C in invasive breast cancer correlates with a poorer response to anti-hormone therapy and reduced patient survival. *J Pathol* **217**, 32-41 (2009).
86. Roe, J.S. *et al.* Enhancer Reprogramming Promotes Pancreatic Cancer Metastasis. *Cell* **170**, 875-888 e20 (2017).
87. Clark, K.L., Halay, E.D., Lai, E. & Burley, S.K. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* **364**, 412-20 (1993).

88. Soufi, A. *et al.* Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161**, 555-568 (2015).
89. Fernandez Garcia, M. *et al.* Structural Features of Transcription Factors Associating with Nucleosome Binding. *Mol Cell* **75**, 921-932 e6 (2019).
90. Roberts, G.A. *et al.* Dissecting OCT4 defines the role of nucleosome binding in pluripotency. *Nat Cell Biol* **23**, 834-845 (2021).
91. Iwafuchi, M. *et al.* Gene network transitions in embryos depend upon interactions between a pioneer transcription factor and core histones. *Nat Genet* **52**, 418-427 (2020).
92. Donovan, B.T. *et al.* Basic helix-loop-helix pioneer factors interact with the histone octamer to invade nucleosomes and generate nucleosome-depleted regions. *Mol Cell* **83**, 1251-1263 e6 (2023).
93. Yan, C., Chen, H. & Bai, L. Systematic Study of Nucleosome-Displacing Factors in Budding Yeast. *Mol Cell* **71**, 294-305 e4 (2018).
94. Feng, X.A. *et al.* GAGA Factor Overcomes 1D Diffusion Barrier by 3D Diffusion in Search of Nucleosomal Targets. *bioRxiv* (2023).
95. Liu, H. *et al.* Visualizing long-term single-molecule dynamics in vivo by stochastic protein labeling. *Proc Natl Acad Sci U S A* **115**, 343-348 (2018).
96. Sekiya, T., Muthurajan, U.M., Luger, K., Tulin, A.V. & Zaret, K.S. Nucleosome-binding affinity as a primary determinant of the nuclear mobility of the pioneer transcription factor FoxA. *Genes & development* **23**, 804-809 (2009).
97. Li, G. & Widom, J. Nucleosomes facilitate their own invasion. *Nat Struct Mol Biol* **11**, 763-9 (2004).
98. Tims, H.S., Gurunathan, K., Levitus, M. & Widom, J. Dynamics of nucleosome invasion by DNA binding proteins. *J Mol Biol* **411**, 430-48 (2011).
99. Polach, K.J. & Widom, J. Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation. *J Mol Biol* **254**, 130-49 (1995).
100. Tan, Z.Y. *et al.* Heterogeneous non-canonical nucleosomes predominate in yeast cells in situ. *Elife* **12**(2023).
101. Liu, Y. *et al.* Widespread Mitotic Bookmarking by Histone Marks and Transcription Factors in Pluripotent Stem Cells. *Cell reports (Cambridge)* **19**, 1283-1293 (2017).
102. Poirier, M.G., Bussiek, M., Langowski, J. & Widom, J. Spontaneous access to DNA target sites in folded chromatin fibers. *J Mol Biol* **379**, 772-86 (2008).
103. Donovan, B.T., Chen, H., Jipa, C., Bai, L. & Poirier, M.G. Dissociation rate compensation mechanism for budding yeast pioneer transcription factors. *Elife* **8**(2019).
104. Koerber, R.T., Rhee, H.S., Jiang, C. & Pugh, B.F. Interaction of transcriptional regulators with specific nucleosomes across the Saccharomyces genome. *Mol Cell* **35**, 889-902 (2009).
105. Yu, X. & Buck, M.J. Defining TP53 pioneering capabilities with competitive nucleosome binding assays. *Genome Res* **29**, 107-115 (2019).
106. Echigoya, K. *et al.* Nucleosome binding by the pioneer transcription factor OCT4. *Sci Rep* **10**, 11832 (2020).
107. Michael, A.K. *et al.* Mechanisms of OCT4-SOX2 motif readout on nucleosomes. *Science* **368**, 1460-1465 (2020).
108. Tanaka, H. *et al.* Interaction of the pioneer transcription factor GATA3 with nucleosomes. *Nat Commun* **11**, 4136 (2020).
109. Zhu, F. *et al.* The interaction landscape between transcription factors and the nucleosome. *Nature (London)* **562**, 76-81 (2018).

110. Li, S., Zheng, E.B., Zhao, L. & Liu, S. Nonreciprocal and Conditional Cooperativity Directs the Pioneer Activity of Pluripotency Transcription Factors. *Cell Rep* **28**, 2689-2703 e4 (2019).
111. Tan, C. & Takada, S. Nucleosome allostery in pioneer transcription factor binding. *Proc Natl Acad Sci U S A* **117**, 20586-20596 (2020).
112. Mivelaz, M. *et al.* Chromatin Fiber Invasion and Nucleosome Displacement by the Rap1 Transcription Factor. *Mol Cell* **77**, 488-500 e9 (2020).
113. Hansen, J.L. & Cohen, B.A. A quantitative metric of pioneer activity reveals that HNF4A has stronger in vivo pioneer activity than FOXA1. *Genome Biology* **23**, 221-221 (2022).
114. Donaghey, J. *et al.* Genetic determinants and epigenetic effects of pioneer-factor occupancy. *Nat Genet* **50**, 250-258 (2018).
115. Lupien, M. *et al.* FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132**, 958-70 (2008).
116. Mayran, A. *et al.* Pioneer factor Pax7 deploys a stable enhancer repertoire for specification of cell fate. *Nat Genet* **50**, 259-269 (2018).
117. Nevil, M., Gibson, T.J., Bartolutti, C., Iyengar, A. & Harrison, M.M. Establishment of chromatin accessibility by the conserved transcription factor Grainy head is developmentally regulated. *Development* **147**(2020).
118. Cirillo, L.A. *et al.* Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Mol Cell* **9**, 279-89 (2002).
119. Swinstead, E.E. *et al.* Steroid Receptors Reprogram FoxA1 Occupancy through Dynamic Chromatin Transitions. *Cell* **165**, 593-605 (2016).
120. Geusz, R.J. *et al.* Sequence logic at enhancers governs a dual mechanism of endodermal organ fate induction by FOXA pioneer factors. *Nat Commun* **12**, 6636 (2021).
121. Mayran, A. *et al.* Pioneer and nonpioneer factor cooperation drives lineage specific chromatin opening. *Nat Commun* **10**, 3807 (2019).
122. Cernilogar, F.M. *et al.* Pre-marked chromatin and transcription factor co-binding shape the pioneering activity of Foxa2. *Nucleic Acids Res* **47**, 9069-9086 (2019).
123. Wang, A. *et al.* Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. *Cell Stem Cell* **16**, 386-99 (2015).
124. Taube, J.H., Allton, K., Duncan, S.A., Shen, L. & Barton, M.C. Foxa1 functions as a pioneer transcription factor at transposable elements to activate Afp during differentiation of embryonic stem cells. *Journal of Biological Chemistry* **285**, 16135-16144 (2010).
125. Sérandour, A.A. *et al.* Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers. *Genome research* **21**, 555-565 (2011).
126. Filtz, T.M., Vogel, W.K. & Leid, M. Regulation of transcription factor activity by interconnected post-translational modifications. *Trends Pharmacol Sci* **35**, 76-85 (2014).
127. Gao, S. *et al.* Chromatin binding of FOXA1 is promoted by LSD1-mediated demethylation in prostate cancer. *Nat Genet* **52**, 1011-1017 (2020).
128. Wang, Z. *et al.* SETD7 functions as a transcription repressor in prostate cancer via methylating FOXA1. *Proc Natl Acad Sci U S A* **120**, e2220472120 (2023).
129. Teng, M., Zhou, S., Cai, C., Lupien, M. & He, H.H. Pioneer of prostate cancer: past, present and the future of FOXA1. *Protein Cell* **12**, 29-38 (2021).
130. Barral, A. & Zaret, K.S. Pioneer factors: roles and their regulation in development. *Trends Genet* (2023).
131. Cai, N., Li, M., Qu, J., Liu, G.H. & Izpisua Belmonte, J.C. Post-translational modulation of pluripotency. *J Mol Cell Biol* **4**, 262-5 (2012).
132. Williams, C.A.C., Soufi, A. & Pollard, S.M. Post-translational modification of SOX family proteins: Key biochemical targets in cancer? *Semin Cancer Biol* **67**, 30-38 (2020).

133. Bascunana, V. *et al.* Chromatin opening ability of pioneer factor Pax7 depends on unique isoform and C-terminal domain. *Nucleic Acids Res* **51**, 7254-7268 (2023).
134. Grossman, S.R. *et al.* Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proceedings of the National Academy of Sciences* **114**(2017).
135. Zeigler, R.D. & Cohen, B.A. Discrimination between thermodynamic models of cis-regulation using transcription factor occupancy data. *Nucleic Acids Res* **42**, 2224-34 (2014).
136. Liu, J., Shively, C.A. & Mitra, R.D. Quantitative analysis of transcription factor binding and expression using calling cards reporter arrays. *Nucleic Acids Res* **48**, e50 (2020).
137. Szczesnik, T., Chu, L., Ho, J.W.K. & Sherwood, R.I. A High-Throughput Genome-Integrated Assay Reveals Spatial Dependencies Governing Tcf7l2 Binding. *Cell Syst* **11**, 315-327 e5 (2020).
138. Levo, M. *et al.* Systematic Investigation of Transcription Factor Activity in the Context of Chromatin Using Massively Parallel Binding and Expression Assays. *Mol Cell* **65**, 604-617 e6 (2017).
139. Xu, C. *et al.* Systematic Dissection of Sequence Features Affecting the Binding Specificity of a Pioneer Factor Reveals Binding Synergy Between FOXA1 and AP-1. *bioRxiv* (2023).
140. Vanzan, L. *et al.* High throughput screening identifies SOX2 as a super pioneer factor that inhibits DNA methylation maintenance at its binding sites. *Nat Commun* **12**, 3337 (2021).
141. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res* **22**, 1798-812 (2012).
142. Arvey, A., Agius, P., Noble, W.S. & Leslie, C. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res* **22**, 1723-34 (2012).
143. Spitz, F. & Furlong, E.E. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613-26 (2012).
144. Gordan, R. *et al.* Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep* **3**, 1093-104 (2013).
145. Abe, N. *et al.* Deconvolving the recognition of DNA shape from sequence. *Cell* **161**, 307-18 (2015).
146. Kim, S. *et al.* DNA-guided transcription factor cooperativity shapes face and limb mesenchyme. *bioRxiv* (2023).
147. Kaluscha, S. *et al.* Evidence that direct inhibition of transcription factor binding is the prevailing mode of gene and repeat repression by DNA methylation. *Nat Genet* **54**, 1895-1906 (2022).
148. Neikes, H.K. *et al.* Quantification of absolute transcription factor binding affinities in the native chromatin context using BANC-seq. *Nat Biotechnol* (2023).
149. Brodsky, S., Jana, T. & Barkai, N. Order through disorder: The role of intrinsically disordered regions in transcription factor binding specificity. *Curr Opin Struct Biol* **71**, 110-115 (2021).
150. Slattery, M. *et al.* Absence of a simple code: how transcription factors read the genome. *Trends Biochem Sci* **39**, 381-99 (2014).
151. Keilwagen, J., Posch, S. & Grau, J. Accurate prediction of cell type-specific transcription factor binding. *Genome Biol* **20**, 9 (2019).
152. Luo, Y., North, J.A., Rose, S.D. & Poirier, M.G. Nucleosomes accelerate transcription factor dissociation. *Nucleic Acids Res* **42**, 3017-27 (2014).

153. Guan, R., Lian, T., Zhou, B.R., Wheeler, D. & Bai, Y. Structural mechanism of LIN28B nucleosome targeting by OCT4. *Mol Cell* **83**, 1970-1982 e6 (2023).
154. Balsalobre, A. & Drouin, J. Pioneer factors as master regulators of the epigenome and cell fate. *Nat Rev Mol Cell Biol* **23**, 449-464 (2022).
155. Bulyk, M.L., Drouin, J., Harrison, M.M., Taipale, J. & Zaret, K.S. Pioneer factors - key regulators of chromatin and gene expression. *Nat Rev Genet* (2023).
156. Rossi, M.J. *et al.* A high-resolution protein architecture of the budding yeast genome. *Nature* **592**, 309-314 (2021).
157. Bernardo, G.M. & Keri, R.A. FOXA1: a transcription factor with parallel functions in development and cancer. *Biosci Rep* **32**, 113-30 (2012).
158. Fakhouri, T.H., Stevenson, J., Chisholm, A.D. & Mango, S.E. Dynamic chromatin organization during foregut development mediated by the organ selector gene PHA-4/FoxA. *PLoS Genet* **6**(2010).
159. Serandour, A.A. *et al.* Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers. *Genome Res* **21**, 555-65 (2011).
160. Wang, H. *et al.* A systematic approach identifies FOXA1 as a key factor in the loss of epithelial traits during the epithelial-to-mesenchymal transition in lung cancer. *BMC Genomics* **14**, 680 (2013).
161. Li, J., Zhang, S., Zhu, L. & Ma, S. Role of transcription factor FOXA1 in non-small cell lung cancer. *Mol Med Rep* **17**, 509-521 (2018).
162. Eeckhoute, J., Carroll, J.S., Geistlinger, T.R., Torres-Arzayus, M.I. & Brown, M. A cell-type-specific transcriptional network required for estrogen regulation of cyclin D1 and cell cycle progression in breast cancer. *Genes Dev* **20**, 2513-26 (2006).
163. Sadelain, M., Papapetrou, E.P. & Bushman, F.D. Safe harbours for the integration of new DNA in the human genome. *Nat Rev Cancer* **12**, 51-8 (2011).
164. Slaymaker, I.M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84-8 (2016).
165. Oceguera-Yanez, F. *et al.* Engineering the AAVS1 locus for consistent and scalable transgene expression in human iPSCs and their differentiated derivatives. *Methods* **101**, 43-55 (2016).
166. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through Propagating Activation Differences. *International Conference on Machine Learning, Vol 70* **70**(2017).
167. Lundberg, S.M. & Lee, S.I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems 30 (Nips 2017)* **30**(2017).
168. Shrikumar, A. *et al.* Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5. 6.5. *arXiv preprint arXiv:1811.00416* (2018).
169. Olive, M. *et al.* A dominant negative to activation protein-1 (AP1) that abolishes DNA binding and inhibits oncogenesis. *J Biol Chem* **272**, 18586-94 (1997).
170. Biddie, S.C. *et al.* Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Mol Cell* **43**, 145-55 (2011).
171. Srivastava, D., Aydin, B., Mazzoni, E.O. & Mahony, S. An interpretable bimodal neural network characterizes the sequence and preexisting chromatin predictors of induced transcription factor binding. *Genome Biol* **22**, 20 (2021).
172. Kleinschmidt, H., Xu, C. & Bai, L. Using Synthetic DNA Libraries to Investigate Chromatin and Gene Regulation. *Chromosoma* **132**, 167-189 (2023).

173. Teytelman, L., Thurtle, D.M., Rine, J. & van Oudenaarden, A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc Natl Acad Sci U S A* **110**, 18602-7 (2013).
174. Farley, E.K. *et al.* Suboptimization of developmental enhancers. *Science* **350**, 325-8 (2015).
175. Crocker, J. *et al.* Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell* **160**, 191-203 (2015).
176. Hovland, A.S. *et al.* Pluripotency factors are repurposed to shape the epigenomic landscape of neural crest cells. *Dev Cell* **57**, 2257-2272 e5 (2022).
177. Bejjani, F., Evanno, E., Zibara, K., Piechaczyk, M. & Jariel-Encontre, I. The AP-1 transcriptional complex: Local switch or remote command? *Biochim Biophys Acta Rev Cancer* **1872**, 11-23 (2019).
178. Fu, X. *et al.* FOXA1 upregulation promotes enhancer and transcriptional reprogramming in endocrine-resistant breast cancer. *Proc Natl Acad Sci U S A* **116**, 26823-26834 (2019).
179. Bi, M. *et al.* Enhancer reprogramming driven by high-order assemblies of transcription factors promotes phenotypic plasticity and breast cancer endocrine resistance. *Nat Cell Biol* **22**, 701-715 (2020).
180. Milan, M. *et al.* FOXA2 controls the cis-regulatory networks of pancreatic cancer cells in a differentiation grade-specific manner. *EMBO J* **38**, e102161 (2019).
181. Wolf, B.K. *et al.* Cooperation of chromatin remodeling SWI/SNF complex and pioneer factor AP-1 shapes 3D enhancer landscapes. *Nat Struct Mol Biol* **30**, 10-21 (2023).
182. Vierbuchen, T. *et al.* AP-1 Transcription Factors and the BAF Complex Mediate Signal-Dependent Enhancer Selection. *Mol Cell* **68**, 1067-1082 e12 (2017).
183. Avsec, Z. *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet* **53**, 354-366 (2021).
184. de Almeida, B.P., Reiter, F., Pagani, M. & Stark, A. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nat Genet* **54**, 613-624 (2022).
185. Mirny, L.A. Nucleosome-mediated cooperativity between transcription factors. *Proc Natl Acad Sci U S A* **107**, 22534-9 (2010).
186. Zaret, K.S. Pioneer Transcription Factors Initiating Gene Network Changes. *Annu Rev Genet* **54**, 367-385 (2020).
187. Whitton, H. *et al.* Changes at the nuclear lamina alter binding of pioneer factor Foxa2 in aged liver. *Aging Cell* **17**, e12742 (2018).
188. Paakinaho, V., Swinstead, E.E., Presman, D.M., Grontved, L. & Hager, G.L. Meta-analysis of Chromatin Programming by Steroid Receptors. *Cell Rep* **28**, 3523-3534 e2 (2019).
189. Glont, S.E., Chernukhin, I. & Carroll, J.S. Comprehensive Genomic Analysis Reveals that the Pioneering Function of FOXA1 Is Independent of Hormonal Signaling. *Cell Rep* **26**, 2558-2565 e3 (2019).
190. Bulyk, M.L., Drouin, J., Harrison, M.M., Taipale, J. & Zaret, K.S. Pioneer factors - key regulators of chromatin and gene expression. *Nat Rev Genet* **24**, 809-815 (2023).
191. Carroll, J.S. *et al.* Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* **122**, 33-43 (2005).
192. Hurtado, A., Holmes, K.A., Ross-Innes, C.S., Schmidt, D. & Carroll, J.S. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nat Genet* **43**, 27-33 (2011).
193. Stefflova, K. *et al.* Cooperativity and Rapid Evolution of Cobound Transcription Factors in Closely Related Mammals. *Cell* **154**, 530-540 (2013).

194. Chronis, C. *et al.* Cooperative Binding of Transcription Factors Orchestrates Reprogramming. *Cell* **168**, 442-459 e20 (2017).
195. Liu, Z. & Kraus, W.L. Catalytic-Independent Functions of PARP-1 Determine Sox2 Pioneer Activity at Intractable Genomic Loci. *Mol Cell* **65**, 589-603 e9 (2017).
196. Buecker, C. *et al.* Reorganization of enhancer patterns in transition from naive to primed pluripotency. *Cell Stem Cell* **14**, 838-53 (2014).
197. Iwafuchi-Doi, M. & Zaret, K.S. Pioneer transcription factors in cell reprogramming. *Genes Dev* **28**, 2679-92 (2014).
198. Heinz, S. *et al.* Effect of natural genetic variation on enhancer selection and function. *Nature* **503**, 487-92 (2013).
199. Gosselin, D. *et al.* Environment drives selection and function of enhancers controlling tissue-specific macrophage identities. *Cell* **159**, 1327-40 (2014).
200. Link, V.M. *et al.* Analysis of Genetically Diverse Macrophages Reveals Local and Domain-wide Mechanisms that Control Transcription Factor Binding and Function. *Cell* **173**, 1796-1809 e17 (2018).
201. Lee, K. *et al.* FOXA2 Is Required for Enhancer Priming during Pancreatic Differentiation. *Cell Rep* **28**, 382-393 e7 (2019).
202. Genga, R.M.J. *et al.* Single-Cell RNA-Sequencing-Based CRISPRi Screening Resolves Molecular Drivers of Early Human Endoderm Development. *Cell Rep* **27**, 708-718 e10 (2019).
203. Lupien, M. *et al.* FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-Specific Transcription. *Cell* **132**, 958-970 (2008).
204. Carroll, J.S., Hurtado, A., Holmes, K.A., Ross-Innes, C.S. & Schmidt, D. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nature genetics* **43**, 27-33 (2011).
205. Chen, J. *et al.* Hierarchical Oct4 Binding in Concert with Primed Epigenetic Rearrangements during Somatic Cell Reprogramming. *Cell reports (Cambridge)* **14**, 1540-1554 (2016).
206. Chen, J. *et al.* H3K9 methylation is a barrier during somatic cell reprogramming into iPSCs. *Nature genetics* **45**, 34-42 (2013).
207. Matoba, S. *et al.* Embryonic development following somatic cell nuclear transfer impeded by persisting histone methylation. *Cell* **159**, 884-95 (2014).
208. Chung, Y.G. *et al.* Histone Demethylase Expression Enhances Human Somatic Cell Nuclear Transfer Efficiency and Promotes Derivation of Pluripotent Stem Cells. *Cell Stem Cell* **17**, 758-766 (2015).
209. Huang, J. *et al.* BIX-01294 increases pig cloning efficiency by improving epigenetic reprogramming of somatic cell nuclei. *Reproduction* **151**, 39-49 (2016).
210. Liu, X. *et al.* H3K9 demethylase KDM4E is an epigenetic regulator for bovine embryonic development and a defective factor for nuclear reprogramming. *Development (Cambridge)* **145**, dev158261-dev158261 (2018).
211. Ji, D. *et al.* FOXA1 forms biomolecular condensates that unpack condensed chromatin to function as a pioneer factor. *Molecular cell* **84**, 244-260.e7 (2024).
212. Guo, C., Ma, X., Gao, F. & Guo, Y. Off-target effects in CRISPR/Cas9 gene editing. *Front Bioeng Biotechnol* **11**, 1143157 (2023).
213. Zhang, M. *et al.* SHIELD: a platform for high-throughput screening of barrier-type DNA elements in human cells. *Nat Commun* **14**, 5616 (2023).
214. Bessen, J.L. *et al.* High-resolution specificity profiling and off-target prediction for site-specific DNA recombinases. *Nat Commun* **10**, 1937 (2019).

215. Durrant, M.G. *et al.* Systematic discovery of recombinases for efficient integration of large DNA sequences into the human genome. *Nat Biotechnol* **41**, 488-499 (2023).
216. Verkuijl, S.A. & Rots, M.G. The influence of eukaryotic chromatin state on CRISPR-Cas9 editing efficiencies. *Curr Opin Biotechnol* **55**, 68-73 (2019).
217. Iacovino, M. *et al.* Inducible cassette exchange: a rapid and efficient system enabling conditional gene expression in embryonic stem and primary cells. *Stem Cells* **29**, 1580-8 (2011).
218. Wissink, E.M., Fogarty, E.A. & Grimson, A. High-throughput discovery of post-transcriptional cis-regulatory elements. *BMC Genomics* **17**, 177 (2016).
219. Maricque, B.B., Chaudhari, H.G. & Cohen, B.A. A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. *Nat Biotechnol* (2018).
220. Gilmour, D.S. & Lis, J.T. Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. *Proc Natl Acad Sci U S A* **81**, 4275-9 (1984).
221. Policastro, R.A. & Zentner, G.E. Enzymatic methods for genome-wide profiling of protein binding sites. *Brief Funct Genomics* **17**, 138-145 (2018).
222. Skene, P.J., Henikoff, J.G. & Henikoff, S. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nat Protoc* **13**, 1006-1019 (2018).
223. Meyer, C.A. & Liu, X.S. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat Rev Genet* **15**, 709-21 (2014).
224. Meers, M.P., Bryson, T.D., Henikoff, J.G. & Henikoff, S. Improved CUT&RUN chromatin profiling tools. *Elife* **8**(2019).
225. Kaya-Okur, H.S., Janssens, D.H., Henikoff, J.G., Ahmad, K. & Henikoff, S. Efficient low-cost chromatin profiling with CUT&Tag. *Nat Protoc* **15**, 3264-3283 (2020).
226. Keller, C.A. *et al.* Effects of sheared chromatin length on ChIP-seq quality and sensitivity. *G3 (Bethesda)* **11**(2021).
227. Penkov, D., Zubkova, E. & Parfyonova, Y. Tn5 DNA Transposase in Multi-Omics Research. *Methods Protoc* **6**(2023).
228. Li, N. *et al.* Tn5 Transposase Applied in Genomics Research. *Int J Mol Sci* **21**(2020).
229. Meers, M.P., Janssens, D.H. & Henikoff, S. Pioneer Factor-Nucleosome Binding Events during Differentiation Are Motif Encoded. *Mol Cell* **75**, 562-575 e5 (2019).
230. Yang, Y. *et al.* The pioneer factor SOX9 competes for epigenetic factors to switch stem cell fates. *Nat Cell Biol* **25**, 1185-1195 (2023).
231. Grandi, F.C., Modi, H., Kampman, L. & Corces, M.R. Chromatin accessibility profiling by ATAC-seq. *Nat Protoc* **17**, 1518-1552 (2022).
232. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-8 (2013).
233. Bentsen, M. *et al.* ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat Commun* **11**, 4267 (2020).
234. Nabet, B. *et al.* Rapid and direct control of target protein levels with VHL-recruiting dTAG molecules. *Nat Commun* **11**, 4687 (2020).
235. Aikawa, Y. *et al.* Treatment of arthritis with a selective inhibitor of c-Fos/activator protein-1. *Nat Biotechnol* **26**, 817-23 (2008).
236. Sekiya, T., Muthurajan, U.M., Luger, K., Tulin, A.V. & Zaret, K.S. Nucleosome-binding affinity as a primary determinant of the nuclear mobility of the pioneer transcription factor FoxA. *Genes Dev* **23**, 804-9 (2009).

237. Mullick, A. *et al.* The cumate gene-switch: a system for regulated expression in mammalian cells. *BMC Biotechnol* **6**, 43 (2006).
238. Panne, D., Maniatis, T. & Harrison, S.C. Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon-beta enhancer. *EMBO J* **23**, 4384-93 (2004).
239. O'Reilly, F.J. & Rappsilber, J. Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. *Nat Struct Mol Biol* **25**, 1000-1008 (2018).
240. Makowski, M.M., Willems, E., Jansen, P.W. & Vermeulen, M. Cross-linking immunoprecipitation-MS (xIP-MS): Topological Analysis of Chromatin-associated Protein Complexes Using Single Affinity Purification. *Mol Cell Proteomics* **15**, 854-65 (2016).
241. Vashee, S., Melcher, K., Ding, W.V., Johnston, S.A. & Kodadek, T. Evidence for two modes of cooperative DNA binding in vivo that do not involve direct protein–protein interactions. *Current biology* **8**, 452-458 (1998).
242. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol Cell* **49**, 825-37 (2013).
243. Chen, H., Yan, C., Dhasarathy, A., Kladde, M. & Bai, L. Investigating pioneer factor activity and its coordination with chromatin remodelers using integrated synthetic oligo assay. *STAR Protoc* **4**, 102279 (2023).
244. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-8 (2011).
245. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).
246. McLeay, R.C. & Bailey, T.L. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics* **11**, 165 (2010).
247. Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-5 (2016).
248. Randolph, L.N., Bao, X., Zhou, C. & Lian, X. An all-in-one, Tet-On 3G inducible PiggyBac system for human pluripotent stem cells and derivatives. *Sci Rep* **7**, 1549 (2017).
249. Hazelbaker, D.Z. *et al.* A multiplexed gRNA piggyBac transposon system facilitates efficient induction of CRISPRi and CRISPRa in human pluripotent stem cells. *Sci Rep* **10**, 635 (2020).
250. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890 (2018).
251. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10-12 (2011).
252. Gaspar, J.M. NGmerge: merging paired-end reads via novel empirically-derived models of sequencing errors. *BMC Bioinformatics* **19**, 536 (2018).
253. Vasimuddin, M., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. *2019 Ieee 33rd International Parallel and Distributed Processing Symposium (Ipdps 2019)*, 314-324 (2019).
254. Barnett, D.W., Garrison, E.K., Quinlan, A.R., Stromberg, M.P. & Marth, G.T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691-2 (2011).
255. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
256. Amemiya, H.M., Kundaje, A. & Boyle, A.P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019).

257. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
258. Batut, B., van den Beek, M., Doyle, M.A. & Soranzo, N. RNA-Seq Data Analysis in Galaxy. *Methods Mol Biol* **2284**, 367-392 (2021).
259. Ross-Innes, C.S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389-93 (2012).
260. O'Connor, T., Grant, C.E., Boden, M. & Bailey, T.L. T-Gene: improved target gene prediction. *Bioinformatics* **36**, 3902-3904 (2020).
261. Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* **10**, 1523 (2019).
262. Tan, S. A modular polycistronic expression system for overexpressing protein complexes in Escherichia coli. *Protein Expr Purif* **21**, 224-34 (2001).
263. Tan, S., Kern, R.C. & Selleck, W. The pST44 polycistronic expression system for producing protein complexes in Escherichia coli. *Protein Expr Purif* **40**, 385-95 (2005).
264. Wang, W.M., Lee, A.Y. & Chiang, C.M. One-step affinity tag purification of full-length recombinant human AP-1 complexes from bacterial inclusion bodies using a polycistronic expression system. *Protein Expr Purif* **59**, 144-52 (2008).
265. Ferguson, H.A. & Goodrich, J.A. Expression and purification of recombinant human c-Fos/c-Jun that is highly active in DNA binding and transcriptional activation. *Nucleic Acids Research* **29**, art. no.-e98 (2001).
266. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-40 (2010).
267. Yoney, A., Bai, L., Brivanlou, A.H. & Siggia, E.D. Mechanisms underlying WNT-mediated priming of human embryonic stem cells. *Development* **149**(2022).
268. Weirauch, M.T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443 (2014).
269. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L. & Noble, W.S. Quantifying similarity between motifs. *Genome Biol* **8**, R24 (2007).
270. Chiu, T.P. *et al.* DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics* **32**, 1211-3 (2016).
271. Chiu, T.P., Rao, S., Mann, R.S., Honig, B. & Rohs, R. Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein-DNA binding. *Nucleic Acids Res* **45**, 12565-12576 (2017).

VITA

EDUCATION BACKGROUND

The Pennsylvania State University University Park, PA
PhD. Molecular Cellular and Integrated Biosciences 2024
Fudan University Shanghai, China
B.S. Biological Science 2015

PUBLICATIONS

[1] **Xu, C.***, Kleinschmidt, H.*, Yang, J., Leith, E., Johnson, J., Tan, S., Mahony, S., & Bai, L. (2023). Systematic Dissection of Sequence Features Affecting the Binding Specificity of a Pioneer Factor Reveals Binding Synergy Between FOXA1 and AP-1. (* Contributed equally). ***bioRxiv***, https://doi.org/10.1101/2023.11.08.566246. **Reviewed in *Molecular Cell*.**

[2] Kleinschmidt, H., **Xu, C.**, & Bai, L. (2023). Using Synthetic DNA Libraries to Investigate Chromatin and Gene Regulation. ***Chromosoma***, 1-23.
[3] Cheng, X., **Xu, C.**, & DeGiorgio, M. (2017). Fast and robust detection of ancestral selective sweeps. ***Molecular ecology***, 26(24), 6871-6891.

AWARDS AND FUNDINGS

Best Student Talk Award at the 39th Penn State Summer Symposium in Molecular Biology (2023)
Huck Graduate Research Innovation Award (2018)
Hui-Chun Chin and Tsung-Dao Lee Chinese Undergraduate Research Endowment (CURE) (2014-2015)
National Top Talented Undergraduate Training Program (NTTUTP) (2013-2015)
Shanghai Scholarship (2014)
National Scholarship (Twice, 2012 & 2013)

CONFERENCE PRESENTATIONS

Oral presentations

Xu, C., & Bai, L. "Systematic dissection of sequence features affecting binding specificity of a pioneer factor". The 39th Penn State Summer Symposium in Molecular Biology, Penn State University, University Park, PA. August, 2023.
Xu, C., & Bai, L. "Systematic dissection of sequence features affecting binding specificity of a pioneer factor". 6th Annual Life Sciences Symposium, Penn State University, University Park, PA. May, 2022.