Genetic Determinants and Epigenetic Effects of Pioneer Factor Occupancy

Genetic determinants and epigenetic effects of pioneer factor occupancy

A dissertation presented

by

Julie Donaghey

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

June 2017

Dissertation Advisor: Alexander Meissner Julie Donaghey

Genetic determinants and epigenetic effects of pioneer factor occupancy

Abstract

Transcription factors (TFs) are the core drivers of gene regulatory networks that control developmental transitions and a complete understanding of how they access, alter and maintain specific gene expression patterns remains an important goal. To begin a systematic dissection of the molecular components that either enable or constrain TF activity, we investigated the genomic occupancy of a set of previously defined pioneer factors, FOXA2, GATA4 and OCT4 in both endogenous and ectopic settings. We find that all three factors display cell type specific occupancy even with super-physiological expression conditions, but only FOXA2 and GATA4 display, in both endogenous and ectopic conditions, low enrichment sampling of additional loci that are occupied in alternative lineages. Ectopic co-expression of FOXA2 and GATA4 can stabilize sites that were previously only sampled. In general, we observe little influence of the chromatin state on FOXA2 or GATA4 enrichment, but a bias towards open chromatin for ectopic OCT4 targets. Finally, we demonstrate that FOXA2 occupancy and changes to DNA accessibility at silent *cis*-regulatory elements can occur when the cell cycle is halted in G1, but surprisingly, subsequent changes in DNA methylation require DNA replication. Taken together, our results provide several new molecular insights that contribute to our basic understanding of gene regulation and pave the way for a more rational use of ectopic TFs for cellular reprogramming.

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Appendix

List of abbreviations

5caC - 5-carboxylcytosine
5hmC – 5-hydroxymethyl-cytosine
5fC – 5-formylcytosine
5mC – 5-methyl-cytosine
bHLH – basoc helix-loop-helix
bp basepair
BER – base excision repair
CGI CPG island
ChIP-BS-seq – chromatin immunoprecipitation bisulfite sequencing
ChIP-seq – chromatin immunoprecipitation
CpG cytosine followed by guanine
dEC – definitive ectoderm
dEN – definitive endoderm
DMR differentially methylated region
dMS – definitive mesoderm
DNAme – DNA methylation
DNMT1 DNA methyltransferase 1
ER – estrogen receptor
FPKM fragments per kilobase of transcript per million mapped reads
GR – glucocorticoid receptor
GRE – glucocorticoid responsive element
H3K27ac histone 3 lysine 27 acetylation
H3K27me3 histone 3 lysine 27 trimethylation
H3K4me1 histone 3 lysine 4 monomethylation
H3K4me2 histone 3 lysine 4 dimethylation
H3K4me3 histone 3 lysine 4 trimethylation
H3K9me3 histone 3 lysine 9 trimethylation
human ESC(s) – human embryonic stem cell (s)
HMR highly methylated region (61-100%)
IMR intermediately methylated region (11-60%)
iPSC induced pluripotent stem cell
kb kilobase
LMR lowly methylated region (0-10%)
MEF murine embryonic fibroblast
MPRA – Massive parallel reporter assay
NER – Nucleotide excision repair
OSK – Oct4, Sox2, Klf4
PBM – Protein Binding Microarray
PTM – post-translational modification
RNA-Seq rna sequencing
RPKM reads per kilobase per million mapped reads
SELEX - Systematic Evolution of Ligands by Exponential Enrichment
TET ten eleven-translocation protein
TF(s) – Transcription factor (s)

TSS transcription start site
WGBS – whole genome bisulfite sequencing

Statement of collaboration

Julie Donaghey completed the majority of the experiments within this work with help from former Harvard undergraduate Jennifer Chen. Sudhir Thakurela completed the computational analysis of the chromatin and expression data analysis and Jocelyn Charlton completed methylation data analysis with guidance from Julie Donaghey and Alex Meissner. The majority of the interpretations are a result of collaboration between Julie, Sudhir and Jocelyn.

Acknowledgments

I would like to thank my advisor, Alex Meissner, for his support and guidance throughout my graduate school career. He has taught me to be critical and patient.

I would like to thank my former mentor, Mitch Guttman, who hired me as a technician at the Broad Institute, introduced me to Genomics, Stem Cell biology and inspired me with his enthusiasm for science.

Thanks to all my colleagues who also became mentors and friends along the way - particularly to Michael Ziller, Casey Gifford, Davide Cacchiarelli and Zack Smith. Thanks to everyone in the Meissner Lab, especially to my collaborators Jocelyn Charlton and Sudhir Thakurela. Thanks to our Broad team (Elena and Hongcang) for washing the HiSeq for me and letting me constantly steal reagents. And thanks to my stellar Harvard undergraduate collaborator, Jennifer Chen, for her work on this project.

Thanks to my parents, Liz and Paul Donaghey for instilling scientific curiosity in me, and supporting through everything I that I do. Thanks to my siblings and extended family and friends for continuous love and support.

Most importantly, thanks to my husband, Ryan Moore, who always makes my day better, no matter how many times the experiment fails

Chapter 1.
Introduction

Parts of this chapter are submitted for publication elsewhere [1].

1.1 Thesis Summary

Organismal development is orchestrated by selective use and interpretation of identical genetic material in individual cells. During this process, transcription factors (TFs) coordinate protein complexes at the associated promoter and distal enhancer elements to repress previously active loci as well as turn on silent genes (**Figure 1.1**). The generally accepted model assumes that primary access to certain regulatory elements can be restricted by chromatin, which could ensure some spatial and temporal control of gene expression during successive developmental stages [2-5]. In fact, most cell type specific TFs, are indeed constrained by chromatin as access to target sites on nucleosomes requires cooperative binding of TF groups or in conjunction with nucleosome remodelers [2,6,7]. In contrast to most TFs, pioneering TFs have been reported to access their target sites even in nucleosomal DNA [2,6,7]. Once bound, pioneer TFs have been shown to possess an intrinsic (ATP-independent) capability to remodel nucleosomes surrounding their target sites [8], and are proposed to create a permissive environment for the coordinated binding of additional factors [2]. Thus pioneer TFs are considered at the apex of the TF hierarchy and are likely critical towards gaining access to silent and repressed genomic loci.

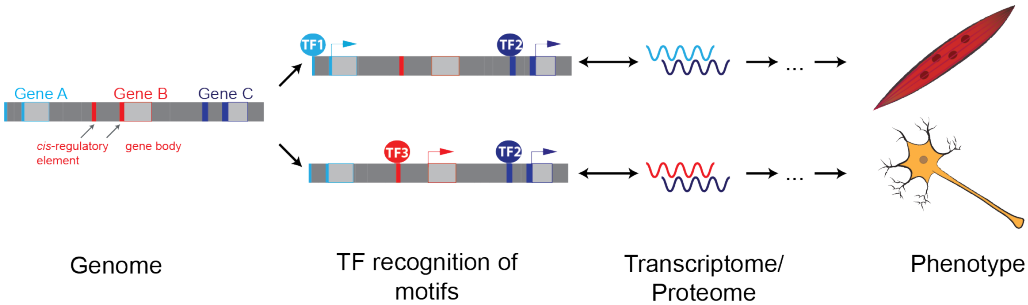


Figure 1.1:
Paradigm illustrating how identical genomic sequences can result in distinct cellular phenotypes by specific TF recognition of cognate motif sequence at previously repressed *cis*-regulatory elements which drives diverse cellular phenotypes.

The focus of my thesis is on FOXA2 which is part of the Forkhead box TF family, that was first characterized as a pioneer factor for its ability to remodel nucleosomes at the repressed enhancers of the *Albumin* locus during endoderm development [6,7,9]. Ablation of FOXA2 in mice is embryonic lethal due to defects in early developmental structures, pointing to a critical role in lineage specification [10,11]. Interestingly however, after early development, FOXA is widely expressed across most endodermal and some ectodermal cell types, suggesting the need for some specificity in its regulation [12,13]. Likewise, studies looking at FOXA1 occupancy across similar breast cancer cell types noted evidence of cell type specific binding [14-16]. Taken together, this suggests that FOXAs specific activity is likely not directed solely by the presence of its cognate DNA motif sequence and that there are perhaps additional features guiding even pioneer factor occupancy.

From these above studies along with other earlier work on pioneer factors, a fundamental question arises which is the focus of the first part of my thesis: How does a factor with supposed universal targeting and remodeling capabilities also exhibit cell type specificity? Initial clues emerged from work examining how cell-type specific co-factors [17-19], signaling [20], and the underlying chromatin landscape [21,22] influence the binding of pioneer factors. However to fully dissect the individual components, it remains a challenge to utilize native developmental systems, where extrinsic signals may induce rapid transitions from initial TF binding to stabilization, local epigenetic remodeling and transcriptional induction without yielding sufficiently stable intermediate states. Genome-wide location analyses within endogenous contexts are subsequently limited to correlations between TF binding, nucleosome occupancy and histone

modifications and cannot distinguish discrete molecular steps. As a result, various studies have attributed differential FOXA/pioneer factor targeting to preferential epigenetic signatures with some inconsistent conclusions [4]. For example, FOXA binding sites when assessed in a steady state have been shown to be enriched in regions of low CpG methylation [14,23], while other studies have suggested that particular FOXA targeted sites only lose CpG methylation after FOXA binding [23-25]. These findings highlight the need for a higher resolution and more systematic study with controllable parameters.

Enhancer elements are critical regulators of gene activation yet our current understanding of the steps needed to coordinate the activation of a silent enhancer elements to ultimately control gene expression are limited (**Figure 1.2)**. We know that repressed enhancer elements are generally nucleosome occluded and contain DNA methylation, and pioneer factors have been shown to remodel nucleosome in vitro [8], yet their abilities to interact with DNAme are unclear. The second part of my thesis focuses on understanding the initial consequences of FOXA2 pioneer factor binding on the surround chromatin structure with an emphasis on understanding the mechanism by which a TF can reprogram DNA methylation patterns upon occupancy.

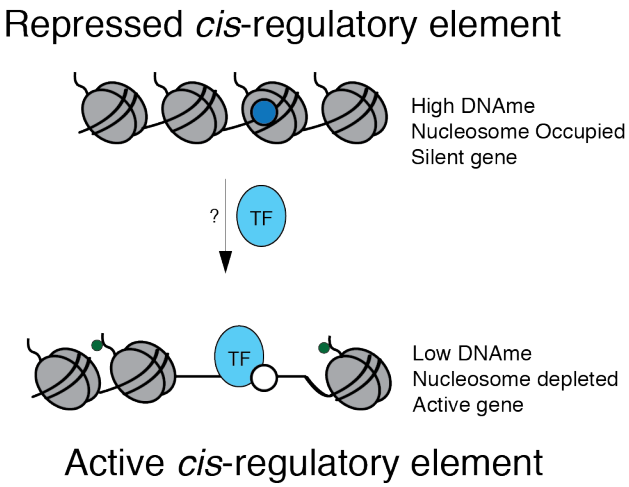


Figure 1.2:
Transition of a repressed *cis*-regulatory element to an active *cis*-regulatory element by a TF. Repressed elements contain high levels of DNA methylation, are nucleosome occluded or have low DNA accessibility and are associated with silent gene expression. When the element transitions to an active state, it contains low levels of DNA methylation, is nucleosome depleted or gains DNA accessibility and is bound by TFs.

From a genome-wide view, we first sought to assess how a pioneer TF gains access to it's target sites by compiling a set of endogenously occupied FOXA2 *cis*-regulatory elements and assessing ectopic FOXA2 targeting at these regions. We engineered an ectopic system, which allowed us to study FOXA2 outside its normal developmental context - removing the factor from an environment that contained other partially redundant FOXA family members as well as known co-factors. This allowed us to systematically dissect the role of distinct cellular components such as cell type specific chromatin architecture and cell type specific transcription factors on FOXA2 occupancy. As not every TF binding site has an immediate measurable output, we next utilized the ectopic system to assess the epigenetic and transcriptional impact of FOXA2 occupancy. Finally, we attempted to gain mechanistic insights into how FOXA2 is able to cause epigenetic remodeling at select closed genomic regions.

Overall, my thesis work begins to assess how the selective targeting of a pioneer factor can initiate events that begin the activation of silent genomic loci. We believe that a detailed molecular understanding of how repressed genomic regions become activated throughout development and reprogramming is a critical step in understanding the molecular basis for cell state conversion. The methods used in this study are broadly applicable for the study of all TFs. Finally, studying the molecular basis of cell state conversion will ultimately lead to better in vitro derived, cellular therapies some day.

In the first part of this introduction, I briefly introduce the key features that influence TF occupancy with a focus on the epigenetic landscape and DNA methylation. I then discuss our current understanding of how TFs can access and occupy target sites in closed chromatin regions. While my thesis focuses on the pioneer factor model for TF binding in closed chromatin, I attempt to compare and contrast the pioneer factor model to the newly emerging dynamic assisted loading/ cooperative TF binding model as assisted loading was recently described as the mechanism by which a small subset of pioneer factor (FOXA) target sites are occupied in breast and prostate cancer cells. Given this, I outline the properties of the two models to better draw conclusions about their distinctions and overlaps. By doing so, it seems that these models may be merging on a similar principle involving TF utilization of ATP-dependent chromatin remodeling machinery. Yet, pioneer factor proteins still hold a distinction in being able to occupy closed chromatin regions without the use of this machinery.

1.2 *cis*-regulatory elements, gene activation and TF binding

The coordinated gene expression in cells is mainly regulated through the interaction of promoter and enhancer regions (i.e. *cis*-regulatory elements) via DNA looping, which function concomitantly to regulate transcription at a particular gene. Specifically, enhancer regions act to coordinate RNA polymerase two along with general transcription factors (pre-initiation complex) to promote gene expression directly from their own DNA sequence as well as to assemble and loop the transcriptional machinery with promoter regions [26,27]. Thus enhancer elements act as the functional units executing proper expression patterning across cells [28]. The human genome encodes tens of thousands of protein coding gene sequences, but an order of magnitude more enhancer elements highlighting the complexity in the regulation of cell type specific gene expression [29]. Multiple enhancers can work together to activate a particular gene and one enhancer can even regulate multiple genes gene forming a complex regulatory network [28].

Active, distal regulatory regions in the genome are often densely occupied by various TFs [30,31]. TFs can interact through direct protein-protein interaction as well as through adjacent DNA interactions as a result of close proximity of motif sequences within regulatory elements. Two main models for TF binding at enhancer regions exist currently [5]. The 'enhancesome' model requires precise, and specific positioning of bound TFs to rapidly induce an 'all-or-nothing' transcriptional response [32]. This type of exact control is actually only found at a subset of genes. Most developmental enhancers follow a 'bill board' model where they can have more flexibility and modulation of TF binding within the regulatory element that does not strictly require all

motifs to be occupied, and can result in more varied levels of transcriptional output [5]. Nevertheless, the coordinated binding and activity of multiple TFs (among other factors) at regulatory elements is a necessary step in activating a silent genomic feature.

TFs play a key role in the precise regulation of gene expression by their ability to specifically recognize and occupy short consensus stretches of DNA sequence, called motifs, across the genome [33]. The recognition sequences of TFs are short, 6-15 nucleotide stretches that are prominent across the genome, yet recent, genome-wide TF mapping studies establish that TFs generally only occupy a small percentage of their possible genomic target sequences [34-36]. Following these studies, we now know that TF occupancy can be influenced by numerous factors, and while the ability to predict TF binding across the genome has increased when considering these influences, predictions are still incomplete. Influences include nucleotide sequence, DNA shape, chromatin and epigenetic modifications, presence of co-binding factors, and three-dimensional architecture of the genome [2,5,37-39].

Most inactive regulatory elements are specifically inaccessible to TF binding yet cell state conversion, either during development or in reprograming, requires activation of previously repressed genes. Thus there is a distinct developmental requirement for a mechanism that transitions repressed regulatory elements to an active state which requires the coordinated binding of regulatory TFs. There are currently two establish methods as to how this might occur. First, there are thought to be a particular unique class of TFs – called pioneer TFs - that can actually access and reorganize repressed chromatin regions by their independent association with target sites in *cis*-regulatory elements. Secondly, in contrast to the independent binding of pioneer TFs, cooperative

binding of TFs along with recruitment of chromatin remodeling machinery – termed dynamic assisted loading - is the second method of how TFs access repressed *cis*-regulatory elements.

1.3 DNA Sequence and Shape

TF binding to DNA is dictated by a combination of the 'on-rate', or the formation of the protein complex on DNA, versus the 'off-rate', or the rate at which said complex dissociates from the DNA sequence [40], as well as a TFs non-specific interaction with the DNA backbone and specific interaction with its motif DNA nucleotide sequences [41]. Numerous *in vitro* methods used to study protein/DNA sequence interactions at high resolution and throughput have given us a better understanding of how specificity of DNA sequence can influence the binding affinity of a particular TF [37,38,40,42-44]. These techniques assess the affinity of a full-length or partial protein against thousands of variable, short DNA pieces [45]. The two main DNA binding assays are:

PBM – (Protein Binding Microarray): Either purified protein or nuclear lysates of cells are washed over a microarray of short DNA sequences to induce an interaction. Fluorescent antibodies are then used to label the protein of interest to highlight, which probes are occupied.

SELEX (Systematic Evolution of Ligands by Exponential Enrichment): Double stranded DNA fragments are washed over an immobilized protein of interest, and the resulting DNA fragments are subsequently removed and sequenced while simultaneously being used as a secondary probe set to evaluate DNA hierarchy.

SELEX increased the length of DNA fragment utilized in the assay improving its accuracy.

In vitro binding studies have been utilized to identify DNA motif preference for many TFs across different cellular contexts. They have revealed that intriguingly, TF family members with similar DNA binding domains can have distinct preferences for core and/or flanking nucleotide sequences despite their structural similarities [39]. Likewise, even individual TFs can have distinct and context specific binding modes due to developmental cofactor relationships in individual cell types [42]. For example, Hox genes prefer a particular DNA sequence motif when in complex with Exd that differs from its independent sequence affinity [38]. Overall these studies revealed that TF interaction with DNA motif sequence is more complicated than originally postulated.

The nucleotide composition surrounding the core TF motif has also been shown to influence TF binding with some TF families having a preference for GC-rich regions while other prefer AT-rich surrounding sequences [37,46,47]. This adjacent sequence likely influences the shape and flexibility of the local DNA adding a structural component that may need to be recognized by the searching TF before occupancy in addition to its motif sequence [37]. Integrating DNA shape features such as GC content, and predicted propeller twist, into TF binding likelihood algorithms can help more accurately predict TF binding at a particular genomic region [37,47]. A recent study demonstrated that subtle preference differences in DNA content surrounding core E-box motifs between two, highly related bHLH TF family members, were predicted to be a result of alterations in DNA shape [46]. While a great deal of information is gained from these studies, utilizing these *in vitro* binding techniques removes TFs from their cellular context and eliminates

the chromatinized context of the DNA sequence which ignores key features TF and DNA interaction.

1.4 Chromatin structure and modifications

DNA does not consist of a simple string of nucleotides and is instead physically condensed by chromatin to allow for the DNA to fit within its nucleus as well as provide protection for the DNA and aide in gene regulation. The fundamental building blocking of chromatin are nucleosomes, which consist of ~147 base pairs of DNA wrapped around an octamer of core histone proteins (2x:H2A, H2B, H3 and H4) [48]. Chromatin can be further compacted about another 20 base pairs by the physical linking of individual nucleosomes by linker histone H1 [49].

Deposition of nucleosomes partially dictates gene expression as active regulatory elements are DNA accessible and repressed genes contain higher order nucleosome structures. DNA accessibility refers to regions that are enriched for signal in DNase hypersensitivity or ATAC-seq studies or regions lacking signal in MNase digestion experiments. The method used to interrogate DNA accessibility has recently become important towards interpreting results, which will be discussed below, in further detail. Overall nucleosome density has been thought to influence the ability of DNA binding factors, mainly TFs and transcriptional machinery, to associate with DNA. Indeed, numerous studies have described the variation in DNA accessibility across distinct cell types and throughout differentiation time courses indicating the dynamic nature of nucleosome positioning is related to cell type specific regulatory regions [50-52].

DNA sequence features, histone variants, post-translational modification to histones proteins influence nucleosome deposition and positioning.

Modulating nucleosome occupancy

First, we know that specific DNA sequences have a higher propensity to be organized into nucleosomes than others [53]. In fact, nearly 50% of all nucleosome positioning is said to be intrinsic organization based on 'bendability' of the DNA sequence [53,54]. Evenly spaced, bendable A/T or T/A dinucleotides assist in nucleotide positioning,[55] while polydA/T or dG/T tracts stiffen DNA and prevent DNA from wrapping around a nucleosome [56]. Eukaryotic promoters often contain stretches of polydA/T making them intrinsically less likely to incorporate nucleosomes [57]. DNA sequence alone however, cannot account for some strongly positioned nucleosomes, such as the +1 nucleosome at DNA promoters, and evidence points to ATP-dependent nucleosome remodeling enzymes along with components of the transcriptional machinery dictating this organization [54,58,59].

While the nucleosome is made up mainly of the canonical histone proteins listed above, variants of these histone proteins are often incorporated into the histone octamer during DNA replication or nucleosome turn-over and can influence nucleosome stability [60]. The histone variant H2A.Z only shares ~60% of its amino acid sequence with its canonical counter part, H2A, giving it distinct structural features [61,62], which have been proposed to weaken the interaction of H2A.Z and the other histone tetramers H3/H4 while also promoting the remodeling activity of ISWI nucleosome remodelers [61,63]. In contrast to the distinct structural changes between H2A.Z and H2A, H3.3 variant

contains only five amino acid substitutions compared to its canonical counter part which result in overall few structural distinctions between the two varieties [64]. Nevertheless, H3.3 deposition occurs independently of DNA replication, in contrast to canonical H3, and incorporates it's self into the nucleosome resourcefully at exposed DNA [65,66]. The distinct variance among both H2A.Z and H3.3 tend to permit their accumulation in areas of active or primed regulatory elements that are likely to have TF engagement [60].

Post-translational modifications (PTMs) of histone proteins that are catalyzed and erased by various histone modifying enzymes (writers and erasers), influence nucleosome occupancy and in turn TF binding. These modifications can directly and indirectly influence the structure of chromatin. For instance, post-translational modifications on the N-terminal tail of all histone proteins can directly influence chromatin structure by chemically destabilizing the interaction between DNA and the nucleosome [67]. Acetylation can neutralize the positive charge of lysine resides, decreasing the affinity interaction between the histone and DNA. Biochemical studies have gone further to demonstrate that incorporation of acetyl groups on the tails of H4 histones, disrupts the 30nm fiber formation inhibiting chromatin compaction [68]. Additionally, indirect modification of the chromatin structure is thought to occur via the histone code hypothesis. This states that distinct PTMs (such as acetylation, methylation ubiquitination, and sumoylation) are subsequently 'read' by specific protein complexes recruited by these modifications, which can, in turn, lead to further modification.

Specific gene regulatory elements are now associated with distinct combinations of post-translational histone modifications and some modifications are associated with

TF binding while others are not [69-78]. First, active promoter regions are associated with mono, di and tri-methylation of histone H3, lysine 4 (ref [73]), as well as acetylation of lysine 9 and 27 (ref [73]). Active enhancer regions contain mono-methylation on lysine 4 and acetylation on lysine 27 of histone H3 (ref [79-85]), while poised enhancer regions are enriched for mono-methylation on lysine 4 (ref [85,86]) plus tri-methylation on lysine 27 of histone H3 (ref [83,84,87,88]). Targeting of these chromatin modifications at regulatory elements is likely dictated in part, by cell-type specific TF binding and also may cause recruitment of cell type specific TFs [89].

MNase titration provides new insights into nucleosome occupancy at active regulatory regions

However, three recent studies report that DNA accessible, active regulatory regions, initially thought to be nucleosome depleted, may in actuality, retain high nucleosome occupancy [3,90,91]. The discrepancies of these new results with decades of previously literature stem from the MNase enzyme used to map nucleosome positioning across the genome. MNase catalyzes the digestion of DNA until it encounters a nucleosome or other obstacle, though results appear to be highly dependent on the enzyme concentration used. These recent studies reveal that comparing both high and low MNase concentrations during digestion conditions can reveal distinct nucleosome structures that are not observed when traditional high MNase concentrations are used alone and highlight the potential variability in nucleosome positioning at regulatory elements [3,90,91]. Many of these active regions that display high nucleosome occupancy under low MNase digestions conditions, and are enriched for nucleosomes that have

high DNA instability, such as histone variants H2A.Z or modified H3K27Ac [90]. Two developmental examples by two different labs provide evidence of this. First, examinations of nucleosome occupancy and DNA accessibility during Unfolded Protein Response (UPR) reveal that there are actually few changes in nucleosome occupancy at regulatory regions of genes that become expressed during this process despite significant gains in DNA accessibility as measured by ATAC-seq [91]. Likewise, FOXA proteins have recently been shown to displace linker histones upon binding 'mnase accessible' nucleosomes which results in gains in DNA accessibility at liver specific, active enhancer regions compared to ubiquitous enhancer elements though the nucleosome remains [3]. These studies are similar to 'fragile nucleosomes' previously identified in yeast and they reveal the highly dynamic nature of nucleosome occupancy at active *cis*-regulatory regions. Further studies are needed to decipher the exact properties of these nucleosomes as they appear to have low stability given they are only identified with very low MNase concentrations. With these papers in mind though, we refer to areas of DNase hypersensitivity/ATAC positive as accessible instead of 'nucleosome depleted'.

1.5 DNA methylation

Methylation of the 5-carbon position on cytosine nucleotides is a covalent modification that allows for added nucleotide variability without any change in genic sequence. Recent investigations into genome-wide DNAme patterns found that global methylation patterns are quite static, with dynamic changes occurring at localized regions and specific genomic features during developmental processes and across cell

types [92]. The methylation of cytosine residues is initially catalyzed mainly by the *de novo* methyltransferase enzymes, DNMT3A/3B, early in development following waves of demethylation after fertilization and during primordial germ cell specification and is subsequently maintained after rounds of DNA replication mainly by the maintenance methyltransferase enzyme, DNMT1. In the following sections I will describe the overall DNAme landscape in somatic cells with a focus on DNAme dynamics at distal *cis*-regulatory elements and how TFs may play a role in the regulation of DNAme at regulatory elements. I will then describe in more detail how DNAme is maintained after replication and present a brief summary of DNA demethylation.

DNAme can be assessed by numerous methods, yet most labs focus on sequencing approaches following bisulfite conversion. Treatment of DNA where sodium bisulfite converts unmethylated cytosine bases to uracil via deamination and then uracil is subsequently converted to thymine during PCR amplification reaction (**Figure 1.3**). This can be done at any scale including the whole genome, which is referred to as whole genome bisulfite sequencing (WGBS). Alternatively, an enzymatically selected set of DNA regions that enriches for DNA with high CpG density referred to as reduced representation bisulfite sequencing (RRBS) can be applied as a more cost effective approach. After sequencing, reads are then mapped back to a reference genome that contains cytosine bases in parallel with a reference genome that converts all cytosine bases to determine methylation status of a cytosine base prior to conversion. A call is generated for each cytosine base in the read and a percent methylation is assigned to each cytosine based on the ratio of methylated to unmethylated reads (**Figure 1.3**).

These assays can be used on whole genomic DNA extractions or from enriched DNA extraction after chromatin immunoprecipitation experiments (ChIP-BS-seq).

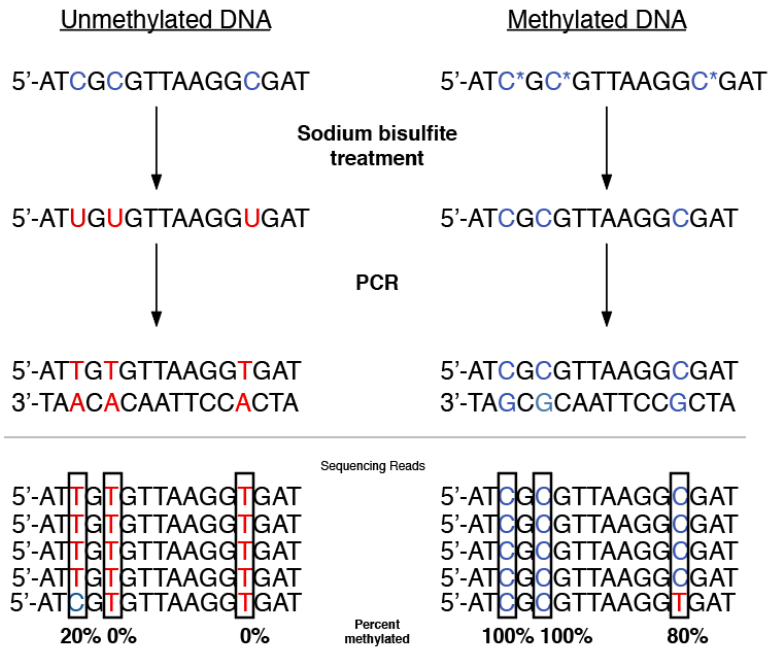


Figure 1.3:
Unmethylated cytosine bases are converted to uracil via deamination, and subsequent PCR amplification then converts the uracil to thymine. Reverse complement strands of DNA then have adenine base paired with the new thymine base. However, methylated cytosine bases remain unconverted and PCR product DNA strands look identical to the DNA prior to bisulfite treatment.

DNAme landscape

The DNAme landscape across the entire genome shows a largely bimodal methylation distribution where CpGs are either highly methylated or completely unmethylated (**Figure 1.4)**. Because spontaneous deamination of 5-methyl-cytosine to thymine in the germline results in the global loss of CpG dinucleotides, the majority of the genome consists of lower than the expected observance rate of CpG dinucleotides (about 28 million CpG total) that are mostly (60-80%) highly methylated [93,94]. However the genome is also punctated by short, CpG dense regions called CpG islands (CGIs)

that contain low levels of DNAme and are generally found at transcription start sites (TSS) regions [95]. CGI promoter regions are generally repressed by processes that are independent of DNAme – mainly by the polycomb repressive complex (PRC) and deposition of the repressive histone modification H3K27me3. Of the ~28 million CpGs in the genome, ~5.5 million CpGs display dynamic changes in DNAme (characterized by at least 30% change in methylation levels) across a large array for cell and tissue types [93]. Regions that display dynamic change in methylation status across cell types are referred to as differentially methylated regions (DMRs) and mainly occur at genomic features distal to the TSS, and fall in areas of the genome that have low, though slightly higher than genome background, CpG density. The majority of annotated DMRs co-localize with DNase I hypersensitivity sites, H3K27Ac enrichment, and/or TF binding clusters in the cell type in which the DMR displays low methylation levels indicating a regulatory function for the majority of these genomic regions as well as the requirement for DNAme loss at active *cis*-regulatory elements prior to TF engagement [93].

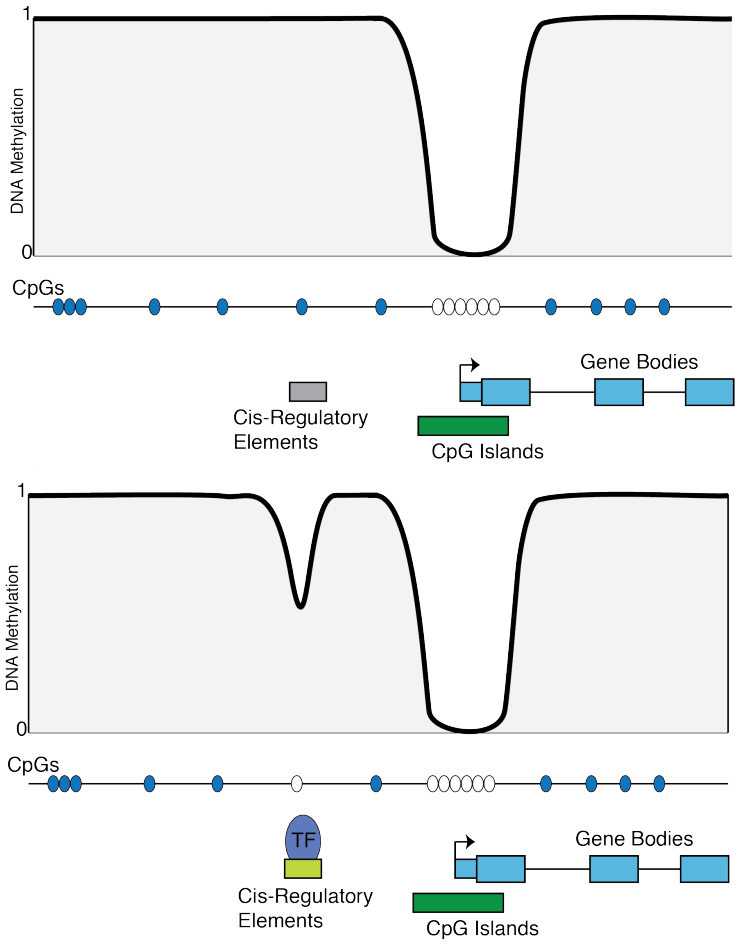


Figure 1.4:
The genome-wide DNA methylation landscape is generally bimodal: genomic features are either highly methylated or completely unmethylated. The genome consists of highly methylated CpGs largely at low density with punctuated regions or high CpG density that are unmethylated. This landscape is overall quite static, yet *cis*-regulatory elements in low CpG density regions undergo dynamic changes in DNA methylation during development and differentiation. When these elements are associated with active gene expression, they are occupied by TFs and contain low levels of DNAme.

DNAme at distal regulatory elements

In particular, during development and differentiation, the local loss of DNAme is observed at regions that gain TF binding and in turn, DNAme is gained at regulatory regions that are no longer functional [25,74,96]. Furthermore, these dynamics are rapid, occurring within hours to days and can have profound regulatory impacts. During a short, five day human embryonic stem cell differentiation into the three germ layers, endoderm (dEN), mesoderm and ectoderm, lineage specific regulatory elements gain DNAme at regions that are occupied by the pluripotency factors (OCT4 and NANOG) them in undifferentiated cells [25]. Interestingly, while gain in DNAme at lineage specific regulatory elements is common in alternative lineages to restrict them, we also observed the gain of DNAme at the FOXA2 CpG island, transcription start site (TSS) during dEN differentiation and in turn, observe the use of a new TSS and an alternative FOXA2 isoform in the dEN state [25].

Our lab showed that during the reprogramming of fibroblasts, embryonic stem cell (ESC) specific enhancer regions that fail to become activated within 96 hours after the onset of reprogramming generally had higher DNA methylation in the starting fibroblast population [97]. Indeed, the addition of DNA methylation inhibitors has been shown to increase reprogramming efficiency [89]. Taken together, this data suggests that DNA methylation in regulatory regions is a repressive mark that needs to be lost before gene activation can occur.

DNAme and TF binding

The relationship between TF binding and DNAme is a complex and the hierarchy is not yet completely understood. There are questions as to whether these active regulatory elements are lowly methylated because of TF binding or if TFs can only occupy these regions due to their unmethylated state [98]. Initially because of the repressive association with DNAme, it was proposed that the presence of DNAme within a TF motif or the surrounding area can actively restrict TF access to those target sites and thus DNAme was thought of as a mechanism for regulating TF binding. There are TFs that are methylation sensitive and are restricted to target sites when they contain DNAme, yet new evidence suggests that DNAme should no longer be considered a general regulatory mechanism for blocking TF/DNA interactions [99,100]. In fact, in recent years, some TF binding events observed at methylated distal regulatory elements are actually also correlated with loss of DNAme [24,25,101] and because of this, people have investigated whether classes of TFs can play a role in the removal of DNAme at regulatory elements. For example, binding of FOXA1 during P19 cell differentiation corresponds with the slow loss of DNAme at enhancer regions and likewise, ChIP-bisulfite experiments demonstrate FOXA2 associated DNA fragments are unmethylated in dEN yet were previously highly methylated in the human ESC starting state – both studies indicating that FOXA may be mediated the loss of DNAme at these target sites [24,25]. Through a transgenic approach by inserting pre-methylated reporter constructs containing a CTCF motif, one group demonstrated that the binding of CTCF protein was not inhibited by pre-existing DNAme and that binding actually correlated with the demethylation of the reporter gene DNA [95]. In contrast though,

examination of CpG rich regions as assessed by RRBS, found CpG methylation at two distinct positions within the CTCF motif are anti-correlated with CTCF occupancy if they contain methylation [102]. Furthermore, IDH mutant gliomas were recently shown to display decreased CTCF binding at functional elements that contain higher methylation levels which results in the loss of CTCF insulator activity and altered gene transcription [103].

These discrepancies in understanding how CTCF is affected by DNAme might be due to the fact that groups are assessing CTCF occupancy at both high and low CpG dense regulatory elements and that CTCF might have the ability to overcome DNAme at low CpG dense regions and occupy those regions, yet its binding can also be methylation sensitive when motifs are found in regulatory elements of high CpG density. Interestingly, when comparing CTCF occupancy in cells that lack DNAme to equivalent cells that contain high levels of DNAme, two groups observed that occupancy of CTCF largely is unaltered across the two cell types indicating that the presence of DNAme at CTCF motifs was likely not obstructing CTCF occupancy [101,104]. The few CTCF binding sites gained in cells lacking DNAme however appear to be distinguished by high CpG content [104]. Thus it seems that DNAme in low CpG areas might be easier to overcome for some TFs than high CpG dense regions.

If TF binding at particular genomic regions can initiate the loss of DNAme, the question remains how and recent studies provide some initial insight. Unlike, CTCF, the methylation-sensitive TF, NRF1, was found to occupy thousands of novel regions in mouse ESCs that have an unmethylated genome compared to wild type conditions [99]. When *de novo* DNA methyltransferases are reintroduced to the system however,

DNAme outcompetes NRF1 for the same regions and novel NRF1 binding sites are lost indicating the binding of NRF1 alone is insufficient to maintain low DNAme levels at target loci [99]. The authors speculate that NRF1 cooperativity with other TFs however may be sufficient to overcome the impending DNAme gain and in fact, find that unmethylated reporter constructs remain unmethylated when NRF1 motifs are flanked by CTCF or RFX motifs but become hypermethylated when the constructs solely contain NRF1 motifs [99]. The authors suggest that either active demethylation would be needed at these NRF1 sites to maintain their unmethylated state or that DNMT enzymes need to be sufficiently blocked by the factor(s) to outcompete *de novo* methylation at these sites.

One could imagine that the ability to occupy and rapidly change the methylation status of CpGs within regulatory regions would be advantageous to pioneer TFs given their potential primary role in locus derepression. Subsequently, the ability to interact with DNAme may be used as a defining pioneering feature and encompasses the third part of this thesis. FOXA (as indicated by the studies above) as well as GATA4 and KLF4 have been previously been characterized for their pioneering closed chromatin binding capabilities as well as their ability to associate with methylated DNA either directly in their core motif sequence or within the surrounding flanking sequence [105,106]. In contrast, reports have shown that the previously characterized pioneer factor OCT4 is unable to bind methylated DNA and generate nucleosome depleted regions when the methylation is within a certain range of its motif, potentially because DNAme restricts movement of the adjacent nucleosome [107,108].

Epigenetic inheritance of DNAme and demethylation mechanisms

To gain insight into how a TF could be mediating the loss of DNAme at particular target sites, it is helpful to first understand more generally how DNAme is maintained through out the cell cycle and the known mechanisms of demethylation. Cytosine methylation mainly exists in the CpG dinucleotide context creating palindromic methylation patterns on both strands of the DNA. Thus this epigenetic information is rapidly copied onto nascent DNA strands after DNA replication mainly due to the affinity of DNMT1 at hemi-methylated DNA. DNMT1 directly interacts with components of the DNA replication machinery, proliferating cell nuclear antigen (PCNA) and UHRF1, which allow for the proper orientation of DNMT1 to methylate nascent DNA strands [92]. The failure to maintain methylation patterns after DNA replication can result in passive loss of DNAme over multiple rounds of division (**Figure 1.5**). In contrast, loss of DNAme is also proposed to occur through the active, enzymatic removal of methylated cytosine bases (discussed in further detail below; **Figure 1.5**).

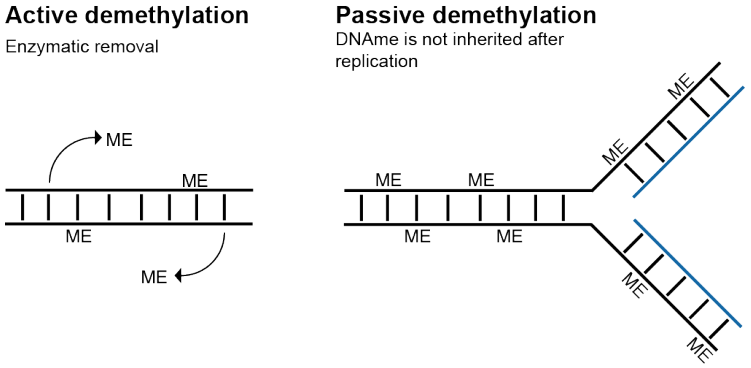


Figure 1.5:
Active demethylation requires the enzymatic removal of methyl groups or methylated cytosine bases directly from the DNA. In contrast, passive demethylation is dependent on DNA replication where nascent DNA strands (shown in blue) fail to retain methyl groups on new cytosine bases.

Notable examples of passive demethylation are pre-implantation development and the specification of primordial germ cells (PGCs) [109]. These precursor cells initially contain somatic levels of (high) DNAme, yet during their migration from the epiblast to the developing gonad, undergo a rapid loss of DNAme globally which ensures the erasure of DNAme from parental imprints, with the exception of some repetitive regions of the genome such as IAPs [110,111]. Quick PGC turnover rates, coupled with precise DNA methylation decay measurements revealed that loss of DNAme is more likely a result of passive depletion following subsequent replication rounds despite the fact that it was initially thought to occur by active demethylation pathways [112,113]. Gene expression profiling data also exposed the down-regulation of a key component in the maintenance methylation machinery complex, *Uhrf1*, during PGC specification [113,114]. While DNMT1 levels remain high during this time period, the loss of UHRF1 which would likely result in significantly impaired nascent strand methylation following replication due to the loss of interaction between DNMT1 and UHRF1[113,114]. It is worth noting that TET1 and TDG expression also remain high during this specification period indicating their potential role in the process [113].

Active demethylation

The active, enzymatic removal of 5mC, that is not dependent on DNA replication processes, is by now an acceptable mechanism for observed loss of DNAme [115]. Active demethylation could result from the following mechanisms:

1. Enzymatic excision of the entire methylated nucleotide by Nucleotide excision repair (NER) or just the base though Base excision repair (BER) pathways. NER

generally functions to remove bulky DNA lesions while BER repairs mismatched or damaged bases[116]. Though there is little evidence to suggest that any of the known mammalian NER proteins function in demethylation [116].

2. Additional modification of the 5-position methyl group via consecutive oxidation reactions or possibly deamination. The modification of 5mC is proposed to impair the recognition of the base by maintenance methylation machinery or allow for a more efficient recognition of the base by repair enzymes.
3. Direct removal of the 5-position methyl group. The active removal of the methyl group at the 5-carbon position would require a demethylating enzyme that can break the carbon-carbon bond critical in this covalent modification and would require high energy requirements [117]. While a few enzymes were initially identified to possess the catalytic capacity to cleave carbon-carbon bonds, none have been implicated in cytosine demethylation [109,116].

A number of enzymes have recently been implicated in active demethylation pathways in mammals. As the direct removal of the methyl group seems unlikely, mounting evidence suggests that a combination of BER mechanisms following modifications to 5mC may be one of the main drivers of this process instead [116]. First, because active demethylation mechanisms were first observed in arabidopsis plants, searches for mammalian homologs initiated the investigations into active demethylation in mammals.

The active, enzymatic excision of 5mC base followed by BER readily occurs in plants and is catalyzed by DNA glycosylases from the Demeter (DME)/repressor of

silencing (ROS) family that cleave the methylated cytosine leaving an abasic and apyrimidine site to be repaired by other enzymes downstream [117,118]. Thymine DNA glycosylase (TDG) in mammals plays a similar role to DME in plants though TDG likely functions after the deamination of 5mC to thymine as its possesses only weak 5mC gylcosylase activity [115]. TDG is required for embryonic development and without it, minimally altered methylation patterns are observed in ESCs upon lineage commitment demonstrating a potential role in methylation regulation [119,120].

The Ten Eleven Translocation (TET) family of enzymes oxidize 5mC to 5-hydroxymethyl cytosine (5hmC) [121] and can then further oxidize 5hmC further to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) [122,123]. Though at much lower levels than 5mC, 5hmC is detected at 1-5% of methylated cytosine bases in most cell types with an increased prevalence in adult neurons [124,125]. It is worth noting that bisulfite conversion of DNA (**Figure 1.2**) cannot distinguish between 5mC and 5hmC. The oxidized varieties of methyl-cytosine bases may be recognized, removed and repaired by similar excision mechanisms mentioned above. Indeed TDG, shows high *in vitro* glycosylase activity on 5fC and 5caC substrates [115]. Additionally, DNMT1/UHRF1 maintenance machinery has a much weaker affinity for hemi-5hMC, which likely limits post-replication methylation activity resulting in passive depletion following oxidation of the cytosine base [126-128].

The oxidation of 5mC to 5hmC by TET enzymes at enhancer regions has been proposed as an initial step in the activation process of the enhancers and is associated with active regulatory elements [101,129,130]. A transition from 5mC to the accumulation of 5hmC is observed at activated distal regulatory elements during neural and adipocyte

differentiation [129]. Furthermore, 5hmC can co-exists at active enhancers with histone modification such as H3K4me2 and H3K27ac and TF occupancy and at promoters of regions of genes that are poised and actively transcribed [129-131]. Increasing 5mC are observed in TET knock out models/knock down experiments which can result in loss of TF binding, active histone modifications, and skewed differentiation abilities likely as a result of delayed gene induction. Although it is worth noting that not all regulatory elements behaved the same way across these experiments indicating varying forms of regulation as well as redundancy in TET proteins.

The question then becomes how can a TF possibly overcome these repressive epigenetic and sequence features to begin the activation process of a repressed *cis*-regulatory element? The remainder of this Introduction attempts to clarify opposing models of how TFs gain access to repressed loci.

1.6 Pioneer factors model

Pioneer factors are a unique class of TF whose exceptional chromatin binding capabilities make them critical regulators of development, reprogramming and cell state transition. Molecularly, pioneer factors have been described to have the following characteristics (**Figure 1.6)**:

1) The ability to occupy target-binding sites that are in closed chromatin as assessed by *in vitro* chromatin array binding assays or by measuring genome-wide binding locations of TFs along with nucleosome positioning or DNA accessibility.

2) The innate ability to remodel nucleosomes, creating a more accessible region following binding. This property was initially predicated on ATP-independent activities of pioneer factors (see below), but has subsequently been less important to the definition.
3) The ability of the factor to bind to target sites before activation of the gene.
4) The critical use of the factor in cell state conversion processes – either reprogramming or development.

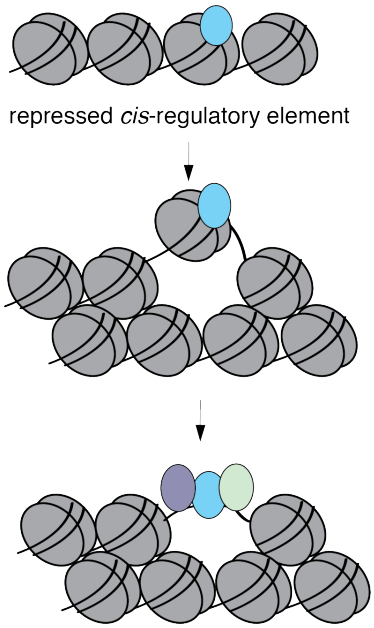


Figure 1.6:
Proposed pioneer factor (blue oval) activation of a nucleosome occluded repressed *cis*-regulatory element which remodels nucleosomes creating more DNA accessibility and presumably allows for the access of additional TFs (purple and green ovals).

In development pioneer factors are thought to establish competence at silent regulatory elements by accessing and remodeling chromatin which allows for subsequent binding by other factors who would not be able to do so independently [2]. This means that pioneer factor binding itself, does not necessary induce a transcriptional response, but rather primes the area for the binding of other factors that

may subsequently induce a transcription response. Because of this, pioneer factors are considered at the top of the TF hierarchy and are thought to be an essential component of the gene activation process.

FOXA TFs can bind and remodel chromatin independently of ATP

The FOXA family of TFs was first characterized as pioneering in the late 1990s for their ability to bind and remodel enhancers of the Albumin locus during endoderm development prior to Albumin expression [132-135]. The family is composed of three structurally similar family members, FOXA1, FOXA2 and FOXA3 and these proteins all contain a DNA binding domain, called a winged helix, which based on crystal structure data, is similar to the DNA binding domain of linker histone H1 [136]. The 'winged helix' domain name is derived from its alpha helix domain and two adjoining winged loop domains that give it a butterfly appearance and allow it to extensively interact, as a monomer, within the major groove of the DNA [137]. The appearance of this domain and suggested that FOXA may have nucleosome remodeling capabilities though its ability to displace linker histones and access one face of the core nucleosome in a similar way to linker histones [138].

This hypothesis was subsequently tested by the Zaret lab by comparing FOXA's remodeling capabilities at H1 compacted versus non-compacted, nucleosome arrays spanning the Albumin enhancer locus [8]. FOXA and GATA easily occupied their binding sites on the non-compacted arrays, though the overall chromatin structure remained intact while other factors, NF-1 and C/EBP known to bind to the same DNA region *in vivo*, could not occupy their sites on the nucleosomal arrays [8]. In contrast, H1

compacted nucleosomal arrays became less compacted upon FOXA, and to a lesser extent GATA occupancy, which suggested that FOXA could physically displace linker histones in an ATP-independent fashion, as these reactions occur in the absence of ATP [8]. Protein deletion experiments revealed that the C-terminal region of the FOXA protein interacts with core histone proteins H3 and H4 and that both the C-terminus and winged helix domain are needed to open compacted nucleosomal arrays [8]. Recently, the Zaret lab also demonstrated that there is an accumulation of linker histone H1 at FOXA-bound liver specific enhancer regions upon the knock out of both FOXA1/2 suggesting the presence of FOXA factors outcompetes linker histone chromatin compaction even *in vivo* [3].

Pioneer factors in reprogramming and development

FOXA proteins are critical regulators of cell state conversion as demonstrated by their critical role in early embryonic development and transdifferentiation protocols. FOXA proteins are expressed early in development (prior to epiblast formation – E4.5) with FOXA2 expression occurring earliest, followed by FOXA1 and FOXA3 expression respectively. Early expression FOXA2 is essential for complete germ layer specification and overall development [139-141] as FOXA2 null mice show an embryonic lethal phenotype due to defects in the node and notochord [142,143]. Most mesodermal cell types formed properly in the FOXA2 null mouse model, which suggested that FOXA2 had less of a role in mesoderm development, although recent studies have demonstrated a role for FOXA during the development of ventricle cardiac cells [144]. Alternatively, FOXA1 null mice, develop normally yet die immediately after birth due to

hypoglycemia and diabetes insipidus demonstrating a role for FOXA proteins in endoderm and liver development [145,146]. Though FOXA1 and FOXA2 are structurally related, the above mouse models suggest that they have distinct regulatory roles in development with FOXA2 being more critical in early cell specification events likely related to its occupancy and chromatin decondensation properties [2]. Furthermore, other mouse models suggest that FOXA2 has a critical role in late endoderm/hepatoblast transition, yet may be dispensable after the initiation of the hepatoblast development. Deleting FOXA2 after hepatoblast formation produces normal mice [147], while deleting FOXA2 during late endoderm development, but pre-hepatoblast formation produces mice that die shortly after birth [148]. Late endoderm cells deficient for both FOXA1 and FOXA2 fail to induce proper liver specification and mice die prior to hepatic bud development [149]. Overall, these mouse models reveal the importance of the FOXA family during developmental state conversions in all three germ layers.

Pioneer TF are also widely used in trans-differentiation protocols along with lineage specific factors to convert one somatic cell type to another somatic cell type [4,150]. For instance, a FOXA factor in combination with HNF4a and HNF1b is sufficient to reprogram a fibroblast into a hepatocyte-like cell [151]. Likewise, as mentioned above, GATA factors, specifically GATA4 is a pioneer factor implicated in the endoderm/hepatic state transitions as well as cardiomyocyte transitions and GATA1/2 are used for hematopoietic transitions [150]. Similarly, the most well-known reprogramming transition - the conversion of fibroblasts to induced pluripotent stem cells – is catalyzed by ectopic expression of the pioneering factors OCT4, SOX2 KLF4 along with, non-pioneer factor,

cMYC [21,22,152]. Presumably, the potent remodeling activities of pioneer TFs allow for the opening of critical target sites for trans-differentiation that are otherwise occluded by chromatin. The initial accessibility within chromatin may then allow for the co-expressed lineage-specific regulators to access their target sites and induce gene expression changes. Alternatively, these lineage factors may function in combination with pioneer factors to dictate the specific subset of targets selected for remodeling, though this has yet to be fully explored. It is clear that pioneer factors have a potent effect and their co-expressed lineage factors indeed help drive cellular state decision, though the final outcome of many transdifferentiation processes results in immature cells that can undergo further conversion based on extrinsic signals [153].

1.7 Cooperative binding of TFs and chromatin remodeling machinery

In contrast to the simple pioneer model, general cooperative binding among TFs has also been proposed as a mechanism to allow TF binding at previously repressed *cis*-regulatory elements (**Figure 1.7)**. In fact, adjacent motif sequences on DNA at the face of a nucleosome can be subject to binding by multiple factors at once, which has been postulated to dislodge a nucleosome from the target site. While each factor independently would not be able to have the energetic capacity to evict a nucleosome from the target location, the combined binding power of multiple factors is thought to be sufficient to compete with the nucleosome for access to the DNA [154]. Few examples of this type of cooperation exist in the literature and the more broadly agreed upon model for cooperative TF occupancy at repressed elements – termed dynamic assisted

loading - dictates TF recruitment of chromatin remodeling machinery to ensure accessibility of the target site for TF binding.

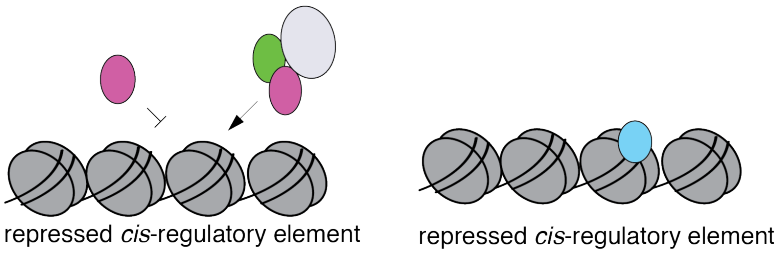


Figure 1.7:
Most TFs cannot access their target sites in repressed chromatin or have to do so cooperatively with other TFs or chromatin remodeling machinery (left schematic). This is in distinct contrast to pioneer factors (right schematic) who have been shown to independently access target sites in repressed chromatin.

Dynamic assisted loading model of enhancer binding

Dynamic assisted loading is defined by the ability of one factor to 'assist' or enhance the loading of a second TF at the same genomic target site, instead of directly competing for binding at that region [155]. This was first observed when examining the binding of the glucocorticoid receptor (GR) and estrogen receptor (ER) at the glucocorticoid responsive element (GRE) via imaging of fluorescently labeled proteins. The authors found that the binding of ER to the GRE is actually enhanced in the presence of GR compared to ER independent expression. Prior to this study, GR had been shown to recruit chromatin remodeling machinery to increase accessibility at some target binding sites and the authors speculated that this recruitment enables GR to remodel the GRE ultimately improving the accessibly of the region for ER binding [156]. When examining occupancy of these factors at GREs in already accessible chromatin, the authors find ER sufficiently occupies these regions and no enhanced binding of ER

is observed after GR co-expression. Instead however, at GREs in pre-existing inaccessible chromatin, the authors observe the enhanced occupancy of ER in the presence of co-expressed GR along with increased DNA accessibility indicating some pioneering capabilities of GR at these regions [155]. It is worth noting that low occupancy of ER exists at GREs located in inaccessible chromatin compared to control even in the absence of GR. Nevertheless, ER occupancy does in fact increase in the presence of GR at the small subset of inaccessible target sites examined. Furthermore, other groups have since described the dynamic assisted loading model similarly for a variety of TFs with a focus on steroid receptor TFs [157-159].

The assisted loading model identifies itself as being particularly distinct from a classical pioneering factor model in three main ways [18].

1. Dynamic assisted loading is based on the idea that multiple factors could be the pioneering or the secondary factor and that these roles are reversible due to the chromatin state of the target site prior to either factors binding.
2. The model relies on the residence times for most TFs being shorter than previously expected and thus the slow chromatin scanning capabilities proposed of pioneer factors is unlikely as TF binding is more dynamic than originally proposed [160].
3. Finally, the model dictates that chromatin-remodeling machinery is integral to the sites being pioneered and that changes in DNA accessibility are not due to ATP-independent mechanisms [18].

Dynamic loading of pioneer factors

Recently, it was demonstrated that for at least a small number of targeted binding sites, nuclear receptor factors ER and GR, can also assisted the loading of the classical pioneer factor, FOXA1 in breast and prostate cancer cells instead of the classically described model where FOXA1 pioneers the target site to allow ER/GR binding [18]. The group first identifies a small number of target sites that FOXA1 occupies specifically after ER/GR induction that are not bound when FOXA1 is expressed independently. They further demonstrate that this small subset of targeted sites gain accessibility only in the presence of co-bound ER/GR and FOXA and attribute this newly found accessibility to the recruitment of chromatin remodeling machinery at these loci, which would indicate a dynamic assisted loading model rather than a pioneering model of TF occupancy.

However, the authors do not present evidence that chromatin-remodeling machinery is recruited to these target sites. Instead, they infer this is the mechanism of action because steroid receptors have previously been shown to recruit remodelers to target sites. Furthermore, a minimal gain in accessibility is observed at target sites where FOXA1 binds independently. Given this information, it is difficult to decipher the cause and effect of the chromatin remodeling at this small set of targeted regions. While the authors assessed DNA accessibility by DNase-seq, it might be helpful to map nucleosomal positioning at these sites as phased nucleosomes, rather than overall increased in accessibility would indicate the presence of chromatin remodelers. Besides, using single molecule tracking studies, the authors find the FOXA1 has similar residence times on DNA to GR and ER - not longer residences times, which was

previously described for pioneer factors though using lower resolution techniques [160]. Given the fast residence times for all factors associated with these DNA regions, cause and effect of chromatin remodeling are even more difficult to ascertain.

Pioneer factors and chromatin remodelers

The main distinguishing feature between the pioneer factor model and dynamic assisted loading model is based in the distinction between who is initiating the change in DNA accessibility at repressed regulatory regions. Is the pioneer factor able to initiate the local chromatin remodeling itself initially or is the factor recruiting ATP-dependent chromatin remodeling machinery simultaneously? When the pioneer factor model was first set forth for the FOXA family members, it proposed that these factors could remodel chromatin independent of ATP remodeling machinery [8] (see above) and it was locally demonstrated *in vitro*. Since this finding though, more pioneer TFs have been linked with chromatin remodeling machinery and it is more accepted that in addition to their ATP-independent features, pioneers might also recruit remodelers to target regions.

Chromatin remodeling complexes hydrolyze ATP to reposition nucleosomes, which overall increases DNA accessibility [161]. While there are four major classes of ATP dependent chromatin remodelers, each family contains similar domain features consisting of an ATPase subunit with distinct auxiliary domains that confer genomic specificity and modulate enzymatic activity [161]. The four major classes are SWI/SNF (BAFs), ISWI, INO80 and CHD,[161] and each of these complexes has distinct roles in chromatin remodeling during development.

FOXA pioneer factors interact with chromatin remodelers *in vivo*, yet whether this interaction is required for the initiation of DNA accessibility changes or to simply stabilize newly accessible regions, remains unclear. During mouse ESC differentiation to the endoderm/hepatic progenitor lineage, regions that become more accessible were bound by FOXA2 in endoderm, but also found to be pre-marked with histone H2A.Z modified nucleosomes in the ESC state [162]. Because INO80/SWIR was shown to mediate the exchange of unmodified H2A histone to H2A.Z [163], the authors examine if chromatin remodelers co- localized at H2A.Z/FOXA2 target site. Indeed, the authors demonstrate that a number of chromatin remodelers accumulate at nucleosome depleted regions that are co-bound by FOXA2 and H2A.Z, [162] and NAP1l1, KAT5 and SMARCA4 were demonstrated to interact with FOXA2 and H2A.Z in endoderm cells by co-immunoprecipitation experiments [128]. *Nap1l1* functions as a cofactor in nucleosome assembly/disassembly, while *Kat5* and *Smarca4* are components of SWI/SNF and SWR1 respectively. While these experiments demonstrate that these factors certainly interact at target loci in the endoderm, with the time resolution of this study, it is difficult to discern if the nucleosome depletion observed is initiated first by FOXA binding or the co-binding of recruited chromatin remodelers as these loci are already functional regulatory elements in endoderm. It is plausible that these factors are recruited to the locus after FOXA binding and nucleosome depletion occurs subsequently to stabilize and even expand the accessible state. Though an intermediate time point was collected and observed, rapid transitions during differentiation system make eliciting these types of cause and effect difficult. Nevertheless, this study highlights that the

classical pioneer factor FOXA has been shown to co-localize and interact with chromatin remodelers and nucleosome disassemblers during development.

GATA proteins have also been shown to interact with chromatin remodeling enzymes through the transactivation domain at the N-terminus of the GATA3 protein and loss of this interaction results in the diminished ability of GATA3 to alter accessibility at a subset of targets [164]. Co-immunoprecipitation studies demonstrated that full length GATA3 and BRG1 protein interact in MDA-MB-231 nuclear extracts [164]. While GATA3 N-terminus deletion constructs can occupy similar genomic regions, these regions fail to elicit a dramatic change in DNA accessibility and BRG1 recruitment to these regions is diminished [164]. This suggests that the n-terminal domain of GATA3 can interact with and possibly recruit BRG1 to particular genomic regions. Though it is worth noting that there are likely other factors contributing to BRG1 recruitment as full length GATA3 protein only recruits BRG1 to a subset of its occupied regions, not every genomic region in which it is bound [164].

Correlations in binding between the pioneering pluripotency factors (OCT4 and SOX2) and certain chromatin remodeler have also recently been demonstrated in ESCs and during reprogramming. The chromatin remodelers CHD1 and INO80, are critical for self-renewal and maintenance of the pluripotent cell state in mouse ESCs as knockdowns lead to decreased expression of pluripotency markers and loss of ESC morphology [165] and *Ino80* is specifically downregulated upon ESC differentiation and pluripotency exit [166]. Furthermore, INO80 is colocalized with the pluripotency master TFs, OCT4, SOX2 and NANOG in ESCs [166]. OCT4 and INO80 have been shown to interact through co-immunoprecipitation studies [166,167] and loss of INO80 is observed at

co-bound regions when *Oct4* is downregulated [166]. Yet, not all OCT4 target sites are also occupied by INO80 which mainly occupies promoter regions compared to OCT4 which occupies both promoter and enhancer regions indicating there are likely other factors involved in INO80 recruitment. Regardless, as these remodelers seem critical for the pluripotent state, there was speculation that CHD1 and INO80 would be required for the establishment of pluripotency during reprogramming of fibroblasts to iPSCs. *Ino80* expression increases concomitantly with the pluripotency factors during reprogramming [168] and INO80 knock down during reprogramming or in embryos leads to decreased number of positive induced PSC colonies [165] and decreased number of viable blastocyst staged embryos respectively [166].

While these papers demonstrate a clear role for chromatin remodeling machinery in the pluripotent state and during reprogramming, they lack explicit evidence that the reprogramming factors themselves, directly recruit chromatin-remodeling complexes to initiate access to target sites in closed chromatin. The deficiency in iPSC reprogramming colonies is some initial evidence that chromatin remodelers may be needed for the establishment of the pluripotent state, and not just the maintenance of the state. Yet these studies do not assess the initial binding events of pluripotency factors in chromatin throughout the reprogramming process and instead only assess the established pluripotent state at the completion of a reprogramming timeline. High resolution ChIP studies that measure the reprogramming factors occupancy at closed chromatin regions along with DNA accessibility in chromatin remodeler deficient MEFs may resolve these timing events.

Considering this mounting evidence, it does appear that pioneer factors may either have the ability to recruit chromatin remodelers themselves or that associated factors somehow recruit remodelers to their targeted regions. Thus it is possible that the pioneer factor model and the assisted loading model are converging on the similar idea: that there is an initiating factor(s) in closed chromatin and subsequent recruitment of remodeling machinery ensures the stability of additional binding events. The fact still remains that pioneer factors themselves, have been demonstrated to access target sites in closed chromatin regions making them unique compared to other TFs and the focus of this thesis work.

1.8 Specific Aims

To decipher how TFs catalyze the initial, molecular events that a repressed *cis*-regulatory element undergoes during activation, first we compiled a set of TF occupied, endogenous *cis*-regulatory to compare and contrast ectopic TF occupancy at using an engineered system of selected TFs, FOXA2, GATA4 and OCT4, with presumed pioneering activities. Despite super-physiological expression for some of the factors, none of them recapitulated high-enrichment binding of their entire, endogenous *cis*-regulatory binding spectrum. However, we provide evidence of a broad, low-level accumulation of FOXA2 and GATA4 signal that might reflect low frequency, 'on-target' sampling at the majority of previously mapped endogenous binding sites; an apparent pioneering feature that appears comparatively distinct from OCT4 which does not display such low-level sampling independently. The underlying epigenetic landscape is informative for factors such as OCT4, but insufficient to explain differential, stabilized

FOXA2 occupancy. Notably, we demonstrate that FOXA enrichment can be partially modulated through cooperation with GATA4 (**Chapter 2**). We observe a wide range of chromatin accessibility dynamics and a general correlation with low enrichment of active histone modifications yet only a fraction of FOXA2 binding sites gained significant DNA accessibility (**Chapter 3**). Finally, we find that FOXA2 can occupy and alter accessibility at chromatinized target regions when the cell cycle is halted in G1, but that epigenetic remodeling of local DNA methylation depends on its presence during DNA replication (**Chapter 4**). From our data a model emerges where FOXA2 uniquely samples the majority of its alternative binding sites, yet it's differential target spectrum is genetically encoded by alternative *cis*-regulatory sequences that are recognized by both FOXA2 and other factors to allow for TF complex stabilization at specific subsets of targets. This model also indicates that occupancy, potentially during S-phase of the cell cycle, may be necessary for epigenetic remodeling of DNAme, but that occupancy at any cell cycle phase examined can induce changes in DNA accessibility. These insights provide a path for more systematic dissection of the functional behavior of different classes of TFs in genomic regulation.

Chapter 2.
Insights into the principles of pioneer factor occupancy

Parts of this chapter are submitted for publication elsewhere [1].

2.1 Rationale

Given the unique binding capabilities of pioneer TFs, one can derive two speculative and opposing models for pioneer factor binding. The first model is based on strict sequence dependence binding, where a pioneer factor can access all its target sites that contain its cis-regulatory motif sequence, regardless of cell state or chromatin environment. This model dictates that binding of a particular pioneer TF will be static across all cell types, and that expression is dictated by the co-occurring lineage specific transcription factors present in the cell. The second model is a context dependent model, where something specific to the cell state (i.e. chromatin landscape, lineage factors, chromatin confirmation etc.) stabilizes pioneer factor binding. In this second model, pioneer factor binding, as well as gene expression changes, will vary across cell types based on context. Mounting evidence has suggested that pioneer factor binding is cell type specific similar to other TFs [16]. In addition, the early embryonic role of many pioneer factors and their expression across vast tissues, suggests the need for some specificity in their binding regulation. With that, the first goal of my thesis was to assess the occupancy of pioneer factor FOXA2 across cell types that endogenously express FOXA2 and compare that to the occupancy of FOXA2 in an ectopic system where FOXA2 is supplied at super-physiological levels and not limited. We next sought to use our ectopic system to gain insights in the rules of FOXA2 occupancy.

2.2 FOXA2 binding at its preferred motif sequence across known regulatory elements

The *FOXA* motif harbors seven core consensus nucleotides along with some less distinct flanking sequence and is therefore, not surprisingly, quite abundant in the human genome [169] (**Appendix S1a).** Because pioneer factors, like the FOXA family, have the unique ability to access target sites in closed chromatin[6,170], one may expect pioneer factors would extensively occupy target loci that contain its core regulatory motif. To specifically assess this we determined what genomic proportion of its preferred motif sequence is actually occupied across a number of human cell types with detectable FOXA2 expression, including HepG2 (hepatocellular liver carcinoma: FPKM 10.9), A549 (lung carcinoma: FPKM 6.2), and embryonic stem cell (ESC) derived definitive endoderm (referred to as dEN [25]; FPKM 20.1). For comparison, we also assessed the proportions of genome-wide motif sequence occupancy CTCF – a factor proposed to have high motif conservation and high residence time on DNA based on its prevalent footprint in DNAse-seq data [171]. We utilized five position weight matrices (PWM) for each factor with varying stringencies (**Appendix S1a**), mapped their positions across the human genome, and then only considered motifs that fell within a region of the genome known to be active across at least one cell type by utilizing all DNAse-seq, H3K27ac, and H3K4me3 data provided by ENCODE (**Appendix S1b**). Altogether we identified around 300,000 motifs across potentially active genome elements of which 6.3-13.7% were significantly bound (utilizing Irreproducibility Discovery Rate (IDR) peak calling on ChIP-seq experiments; see **Methods**) by FOXA2 in any of these cell types (**Figure 2.1a**, **Appendix S1**). Similarly, we find that CTCF

binds to a small percentage of overall motifs (9-12%) despite observing three times as many instances of the CTCF motif across the genome and likewise, about 3 times the number of binding sites across three, matched cell types (**Appendix S1B)**. This finding for CTCF is not surprising though given its cell type specific occupancy [102]. Thus we can conclude that FOXA2, like other DNA binding factors, only occupies a small percentage of its motif sequence across potential active genomic features and its enrichment was largely cell type specific consistent with prior evidence that even pioneer factors display cell type specific binding [14-16] (**Figure 2.1b**). These findings highlight and confirm that the genomic targeting of FOXA2, and possibly other pioneer factors, is not exclusively driven by the presence of its motif sequence alone despite its presumably unrestricted chromatin binding capabilities. Thus even pioneer factor targeting must be influenced by additional factors, which we explore further below.

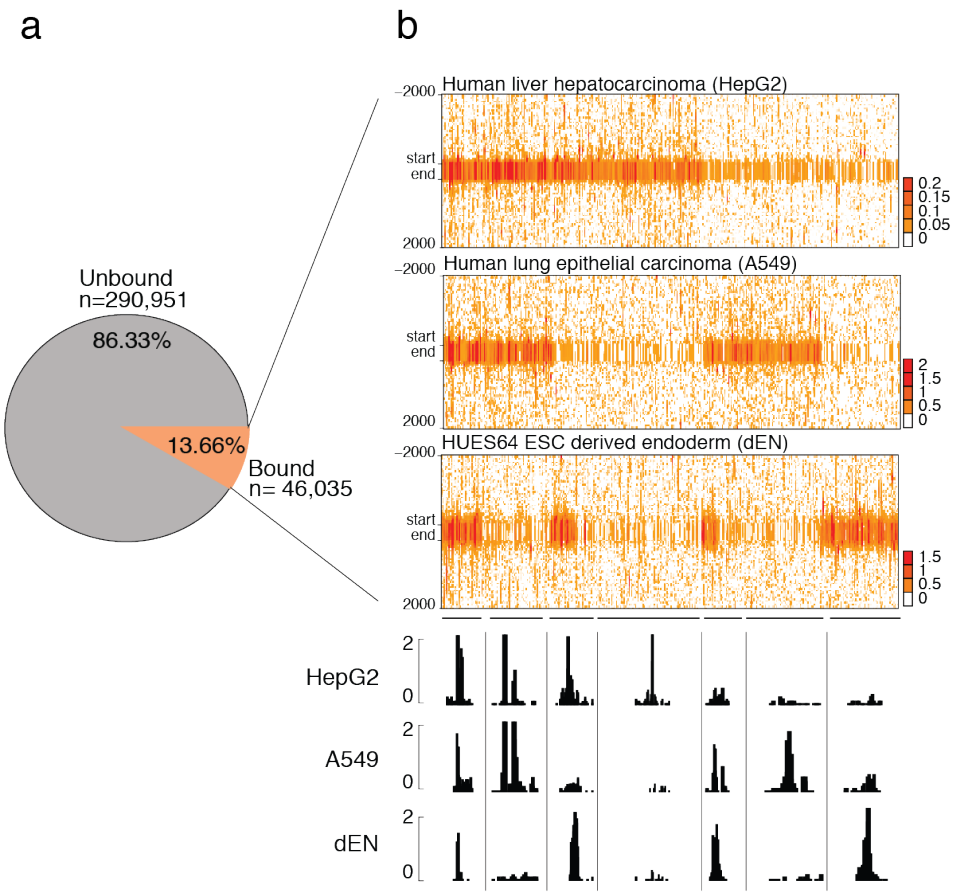


Figure 2.1:
a) Pie chart displays the percentage of FOXA motifs (see Supplementary Fig 1) mapped across the genome that are unbound or bound by FOXA2 at an potentially accessible genomic region in ESC derived endoderm (dEN), HepG2, and A549 cells. b) Read density heat maps for all FOXA2 peaks in dEN, A549 and HepG2 cells that overlap with a motif instance. Heat maps are clustered by occurrence of binding across the three cell types. IGV tracks showing shared genomic occupancy across the three cell types (chr18:9,072,728-9,075,158), unique occupancy in HepG2 (chr18:9,202,880-9,225,100), unique in A549 (chr18:9,008,450-9,022,842) shared occupancy in A549 and HepG2 (chr18:8,725,886-8,734,843) shared in HepG2 and dEN (chr4:80,986,601-81,000,201) shared in A549 and dEN chr4:75,017,694-75,029,960 and unique occupancy in dEN (chr4:74,903,404-74,905,306).

2.3 Ectopic system study of pioneer factor occupancy

Substantial work has been performed to assess the specific regulatory functions of pioneering TF families, yet it remains a challenge to utilize native developmental systems, where extrinsic signals may induce rapid transitions from initial TF binding to

stabilization, local epigenetic remodeling and transcriptional induction without yielding sufficiently stable intermediate states. Genome-wide location analyses within endogenous contexts are subsequently limited to correlations between TF binding, nucleosome occupancy and histone modifications and cannot distinguish discrete molecular steps. These issues highlight the need for a higher resolution and more systematic study with controllable parameters. We therefore engineered a doxycycline (DOX) inducible system in primary foreskin fibroblasts (BJ) that do not normally express FOXA2 or other FOXA family members (**Appendix S2a**). We derived several clonal cell lines with no detectable FOXA2 protein level in the uninduced state, but rapid, uniform and consistent mRNA/protein induction upon DOX treatment (**Figure 2.2, Appendix S2b-d**).

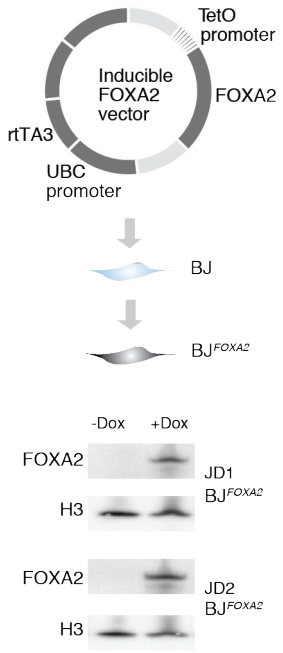


Figure 2.2:
Schematic of the pTripZ vector used for the generation of a clonal FOXA2 inducible cell line BJ^{FOXA2}. Western blot of FOXA2 and H3 protein levels in two, distinct BJ^{FOXA2} clonal lines (JD1 and JD2).

We next performed ChIP-seq for FOXA2 in our BJ^{FOXA2} lines after 1, 4, and 10 days of ectopic induction and found a clear increase in FOXA2 binding sites between 1 and 4 days with little change afterwards (**Appendix S2e-f**). As binding appeared to reach a steady state after 4 days and the overlap between peaks was high, we used our 4 and 10 day time points to identify 49,830 significant consensus IDR peaks. Despite an increase in the total number of FOXA2 peaks in the ectopic system we still primarily observe cell type specific FOXA2 binding showing a limited, statistically significant consensus peak set between the ectopic conditions and any endogenous cell type (~30% consensus FOXA2 between dEN and BJ; **Figure 2.3**). Ectopic FOXA2 binding though does appear in part, driven by DNA motif sequence as the majority of ectopic peaks contain some FOX family motif (region matches: 98.6% [172]; **Appendix S2g**). Yet as the majority of motifs across the genome are unoccupied, this confirms again that DNA sequence is not sufficient for FOXA2 binding. Lastly, even when including our ectopic BJ system, we see no evidence of FOXA2 peak call saturation, suggesting that binding data from additional cell types are expected to confirm more of the FOXA2 motifs as targets (**Appendix S2h**).

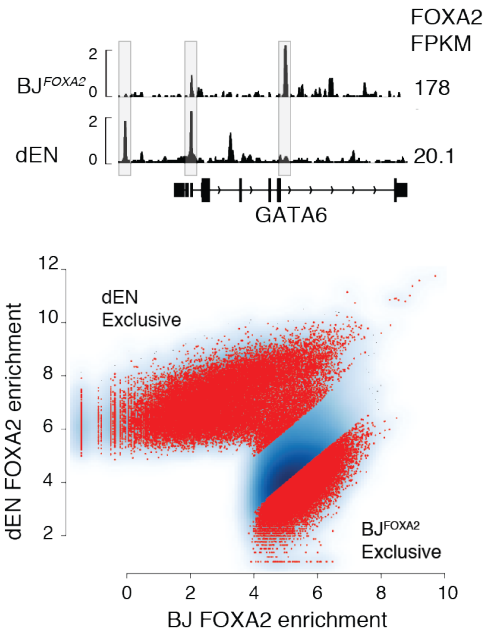


Figure 2.3:
IGV browser shots show differential stabilized binding across ectopic BJ^{FOXA2} and dEN (chr18:19,745,852-19,782,939). FOXA2 FPKMs listed on the right. Scatter plot shows output of DiffBind differential peak set analysis between dEN and BJ^{FOXA2}. Red dots indicate peaks that are considered to be differentially bound between the two data sets.

2.4 FOXA2 and GATA4 display low-level enrichment at the majority of targets in alternative lineages

While we only observed a partial overlap between significantly called peaks in endogenous and ectopic contexts, we nevertheless noticed consistent, low-level FOXA2 enrichment (**see Methods**) at the majority of dEN, HepG2, and A549 bound regions in our ectopic system (**Figure 2.4; left panel; Appendix S4a**). To determine if this low-level enrichment was a general feature of ectopic TF expression or a unique feature of noted pioneer factors, we engineered similar systems in BJ fibroblasts to ectopically express OCT4, and GATA4 (BJ^{OCT}, BJ$^{GATA4;}$ **Appendix S3)**. We observe a similar low-level enrichment pattern in BJ^{GATA4} at GATA4 target sites in dEN and definitive mesoderm (dMS). In contrast, ectopic OCT4 expression in the BJ fibroblasts did not

result in low-level enrichment at regions that are occupied by OCT4 in human ESCs (**Figure 2.4)**. Notably, ectopic OCT4 can display low level enrichment at ESC OCT4 targets when it is co-expressed with SOX2, KLF4 and cMYC in reprogramming BJ fibroblasts [21] indicating its context specific binding differences (**Appendix S4c**). Additionally, by examining FOXA2 enrichment in HepG2 cells and dEN cells at A549 differentially called peaks, we find that this low-level enrichment is also observed in cells endogenously expressing FOXA2 and therefore not just a product of super-physiological levels (**Appendix S4b)**. The additional weaker peaks suggest that FOXA2 and GATA4 may independently sample [17] most of its potential alternative 'on-target' sites but that their binding is mainly enriched at a particular subset of regions defined by additional criteria.

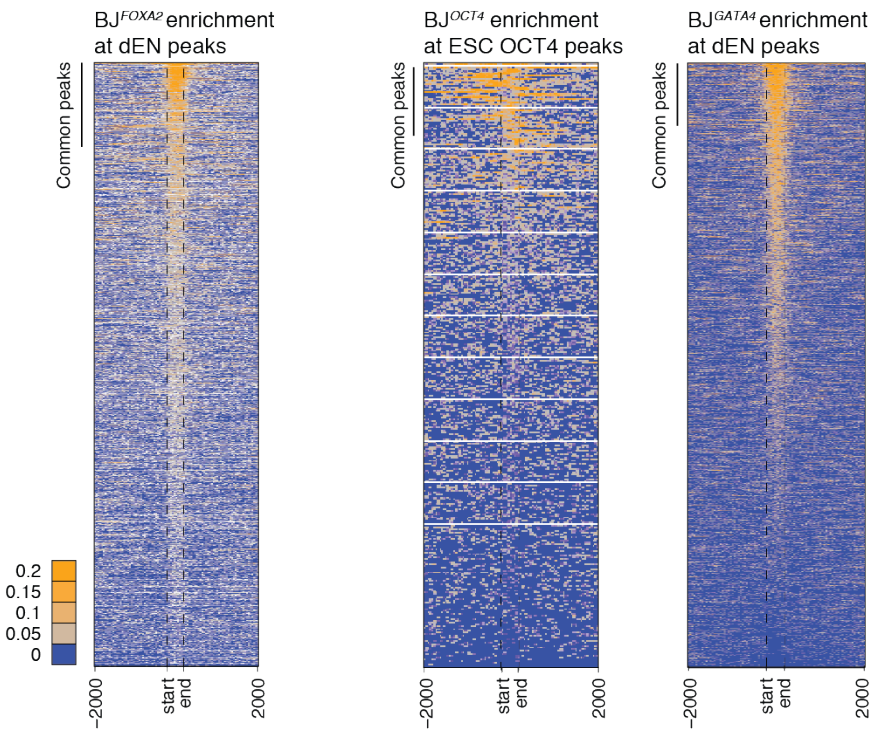


Figure 2.4:
Left: Read density heat maps of FOXA2 enrichment in BJ^{FOXA2} at dEN FOXA2 bound genomic regions. Bar indicates peak calls in common between ectopic FOXA2 ChIP-seq data in BJ^{FOXA2} and dEN FOXA2 ChIP-seq. Dashed lines mark the start and end of FOXA2 peaks with 2kB extension on either side of the peak. Most dEN sites still show low-level enrichment of FOXA2 in induced BJ^{FOXA2} fibroblasts yet are not called as significantly enriched by our MACS peak calling. **Middle:** Read density heat map of OCT4 signal in the BJs at human ESC OCT4 bound genomic regions. Red bar indicates peak calls in common between ectopic OCT4 ChIP-seq data and ESC OCT4 ChIP-seq. In contrast to FOXA2, very few ESC OCT4 sites show any notable level of enrichment of OCT4 binding when ectopically expressed BJ fibroblasts. **Right:** Read density heat map of GATA4 signal in the BJs at human GATA4 dEN bound genomic regions.

To investigate alternative target sampling of FOXA2 further, we plotted the density of FOXA2 enrichment in BJs at the total union set of endogenous and ectopic FOXA2 peaks sets (hence forth referred to as 'FOXA2 union set'). We then compared the peak distribution of FOXA2 to equivalent plots displaying OCT4 and GATA4 enrichment in BJs at the OCT4, GATA4 union sets (referred to as 'OCT4 union set'-

peaks in BJ and ESC, or 'GATA4 union set'- peaks in BJ, dEN and dMS). This analysis clearly shows the distinctive binding properties among these factors with a notable large number of regions displaying intermediate enrichment of FOXA2 and GATA4 compared to more discrete, bimodal distribution for OCT4 (**Figure 2.5**).

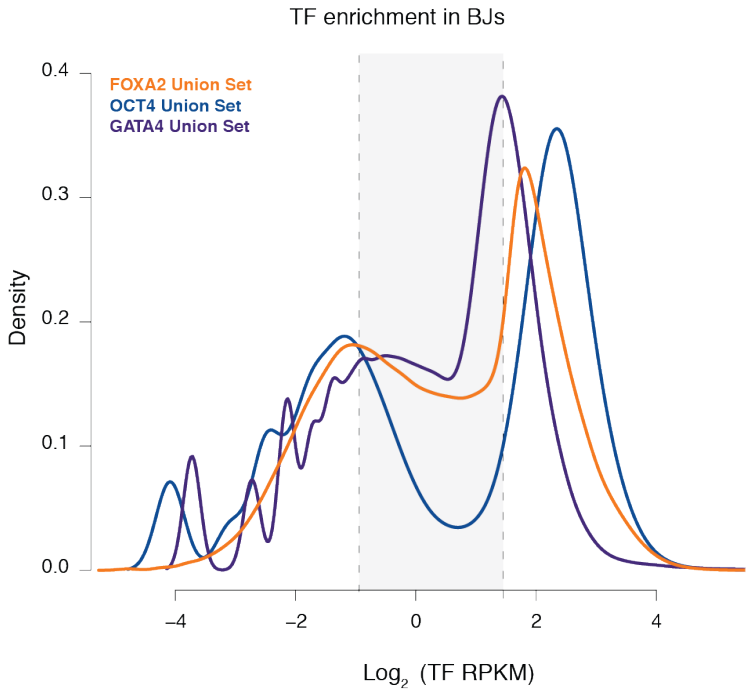


Figure 2.5:
Density histogram displaying FOXA2, OCT4 and GATA4 log2 RPKM enrichment at union sets of ectopic and endogenous sites (FOXA2 – orange, OCT4 – blue, GATA4 - purple). Dashed lines demarcate regions within the background distribution, regions called as sampled sites and regions that were called as peaks.

Overall, we find that FOXA2 and GATA4 exhibit two distinct modes of genomic targeting compared to OCT4. The first mode is characterized by high frequency, highly enriched occupancy at cell type specific targets, while the second mode is characterized by low frequency, lower enriched sites that are alternatively occupied in an another cell type. The term frequency here can apply to both the number of times a TF occupies a target site in a given single cell or how often the site is occupied across a population of cells

as ChIP-seq data cannot distinguish between these possibilities. Regardless, this leads us to speculate that additional nuclear factors assist FOXA2 in cell type specific target sites, which we assess in further detail.

2.5 Differential influence of prior epigenetic state on FOXA2 and GATA4 compared to OCT4 binding

The epigenome is often considered central to establishing and maintaining cell type specific expression patterns, theoretically by restricting access of lineage specific TFs to the DNA. To determine how the cell's pre-existing epigenome may affect FOXA2 binding, we initially investigated the epigenetic landscapes in BJ fibroblasts at BJ^{FOXA2} cell type specific targets. As above, we performed equivalent analyses for OCT4 regions occupied in BJ^{OCT4} fibroblasts and GATA4 regions occupied in BJ^{GATA4} fibroblasts. For these analyses, to define cell type specific FOXA2 targeted regions, we utilized a more stringent cut off than imposed in the differential binding analysis in **Figure 2.3** (see **Methods**) to focus only on the highly enriched targets in a given cell type. To map the epigenetic landscape of our BJ line, we performed ChIP-seq for H3K27ac – which marks transcriptionally-engaged enhancers [173,174], as well as H3K27me3 – a repressive, but reversible, modification, and utilized H3K4me1 ChIP data in NHDF (dermal fibroblasts) cells to examine poised enhancer regions [175]. Additionally we used the assay for transposon-accessible chromatin (ATAC-seq) [176] to map accessible DNA and whole genome bisulfite sequencing (WGBS) to measure DNA methylation. We then defined chromatin states using simple, hierarchical rules that reflect prior knowledge of these modifications and how they interact (**Figure 2.6**). First, 'open' regions were categorized by the occurrence of ATAC-seq enrichment in the pre-

existing BJ chromatin state. Then regions highly enrichment for H3K27ac or H3K4me1 were next categorized as 'active' and 'poised', respectively. Regions enriched for H3K27me3 or H3K9me3 were categorized broadly as 'repressed' and finally all remaining regions that were not classified into one of the above classes were grouped by their DNAme levels: highly methylated regions (HMRs > 60% mean methylation), intermediate methylated regions (IMR mean methylation: 20-60%) and lowly methylated regions (LMR: < 20% mean methylation). LMRs are equivalent to a 'low signal' state that lacks DNA accessibility as well as enrichment of any assessed histone modifications [4].

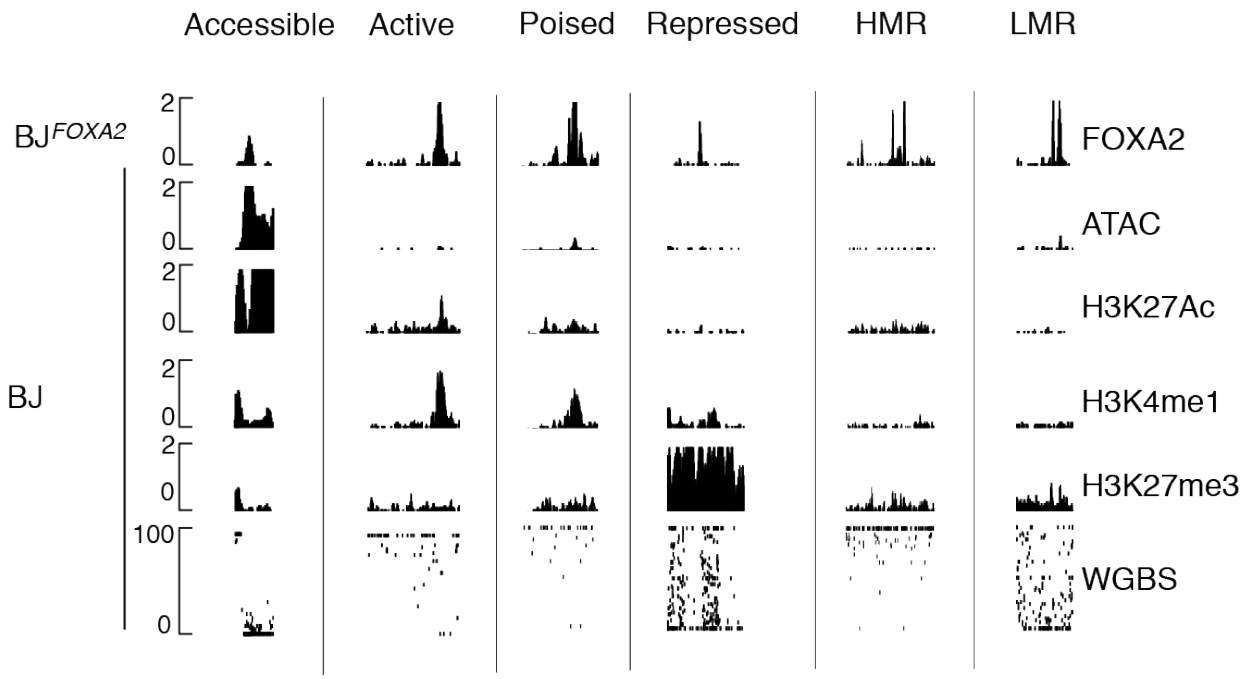


Figure 2.6:
IGV browser tracks displaying FOXA2 binding at each chromatin state we defined (coordinates from left to right: chr6:109,366,481-109,381,042; chr14:75,743,837-75,747,300; chr6:108,485,215-108,512,013; chr6:108,213,193-108,245,723; chr20:43,024,520-43,048,327). Classification was defined and employed hierarchically.

Overall we observe that FOXA2 in the BJs can engage considerable chromatin state diversity with the majority of targets being devoid of our selected histone modifications and instead containing DNAme (**Figure 2.7a**). We find a similar relationship with the epigenome for endogenous FOXA2 targets by assessing the epigenome of ESCs at sites that will be bound in dEN although the background expression of FOXA2 in the undifferentiated state is detectable (**Appendix S5**; FOXA2 ESC 5.6 FPKM and dEN; FPKM 20.1). While GATA4 displays an almost equivalent behavior to FOXA2, we find that a higher percentage of ectopic OCT4 bound sites reside in pre-existing open/active chromatin regions. To compare the epigenomic influence on occupancy of the three TFs in more detail, we used Spearman correlations between TF enrichment and the enrichment of selected epigenetic features (**Figure 2.7b and Appendix S5**). While ectopic FOXA2 and GATA4 enrichment show little positive correlation with any given feature tested, OCT4 binding is highly correlated with accessible pre-existing chromatin (**Figure 2.7b,c**). The majority (55%) of OCT4 ectopic binding sites fall in genomic regions marked by ATAC-seq/active histone modification, while only ~30% of ectopic FOXA2 peaks reside in similar regions (**Figure 2.7a**). The unique OCT4 association with high DNA accessibility is further highlighted by a comparison of scatter plots displaying TF enrichment versus ATAC-seq enrichment (**Figure 2.7c**).

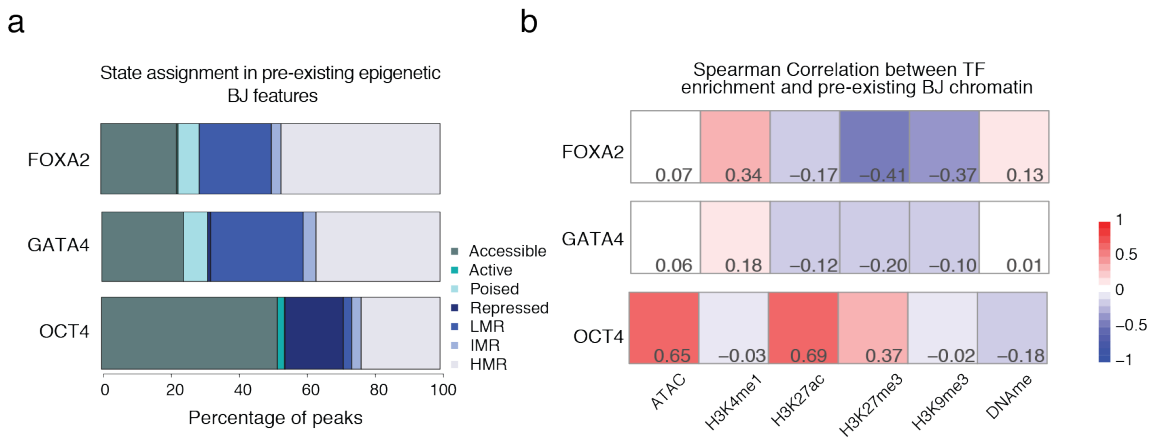


Figure 2.7:
a) Percentage of TF bound regions in BJ^{FOXA2}, BJ^{GATA4}, BJ^{OCT4} falling into assigned chromatin states. State is defined using chromatin in BJ fibroblasts prior to and TF induction. Chromatin state hierarchy is described in **Figure 2.6**. b) Spearman correlations between TF enrichment and epigenetic features displayed as heat map.

We nevertheless observe OCT4 binding sites located in non-accessible DNA regions categorized by their lack in ATAC-seq signal prior to induction (**Figure 2.8**). We find these “closed” regions tend to overlap with lowly enriched H3K27me3 CpG Islands (CGIs) that show canonical depletion of DNAme, which are rarely occupied by FOXA2 and GATA4 (**Figure 2.9, Appendix S5)**. Finally to also compare these behaviors to the ectopic binding of a presumed non-pioneer factor, we generated a similar ectopic BJ system for the hepatocyte nuclear factor, HNF1A and performed ChIP-seq (**Appendix S6).** As may be expected, we were unable to detect significant HNF1A enrichment alone however the factor was readily detectable when co-expressed with FOXA2 (**Appendix S6**) confirming the distinct range in ectopic binding capabilities across pioneer and non-pioneer TFs.

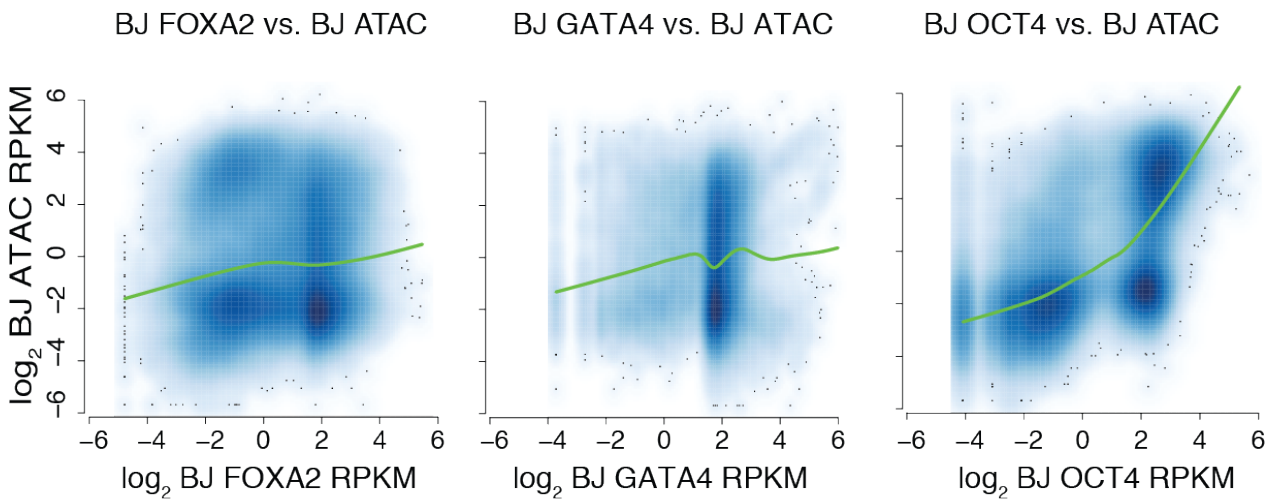


Figure 2.8:
Scatter plots and lowest fit curves (green line) of FOXA2, OCT4 and GATA4 enrichment (Log_2 RPKM) versus ATAC-seq enrichment (Log_2 RPKM).

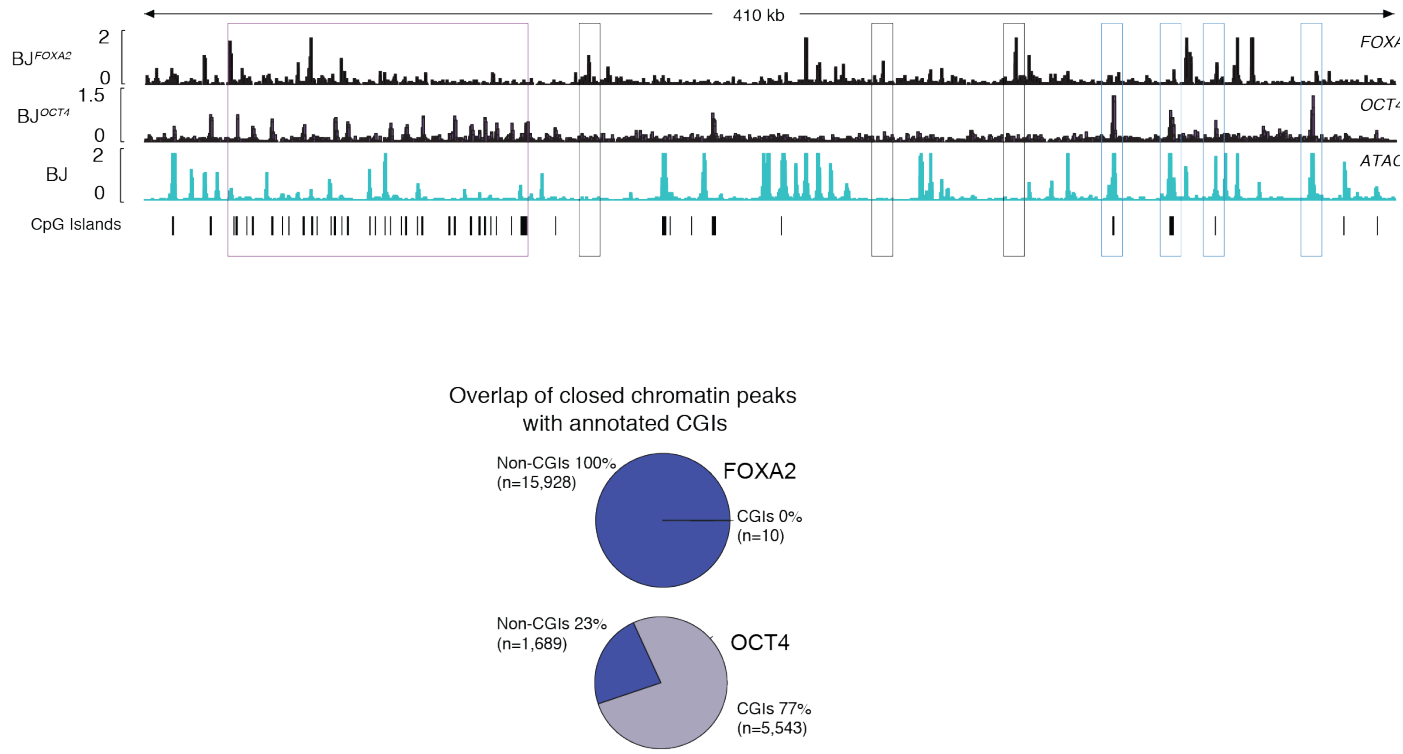


Figure 2.9:
Representative IGV browser tracks displaying FOXA2 and OCT4 enrichment compared to pre-induced BJ ATAC-seq data (chr5:140,657,329-141,085,891). Purple boxed locations highlight regions of OCT4 binding in pre-existing closed chromatin that overlap with annotated CpG islands. Gray boxed locations highlight FOXA2 binding at pre-existing closed chromatin while blue boxed locations highlight OCT4 binding in regions of pre-existing open chromatin. Pie charts to the right of the browser tracks summarize the percentage of FOXA2 and OCT4 regions in pre-existing closed chromatin that overlap with CpG islands.

Lastly, we did notice a minimal anti-correlation of ectopic FOXA2 binding with repressive chromatin modification enrichment, thus we investigated the impact of previously described megabase-scale H3K9me3-marked heterochromatin domains (henceforth referred to as K9-domains; [21], **Appendix S7a-c**) that were reported to restrict OCT4, SOX2, KLF4 (OSK) binding during fibroblast reprogramming to iPSCs [21]. We find that FOXA2 enrichment was generally depleted in these domains with limited FOXA2 binding events (n=417 FOXA2 peaks, with at least 20% overlap of a K9-domain;

mean FOXA2 RPKM: 4; **Appendix S7d)**. Notably, very few exclusive, endogenously occupied FOXA2 regions in HepG2, A549 or dEN cells fall within BJ K9-domains, indicating that these domains are not the major cause of the cell type specific occupancy observed for FOXA2 (**Appendix S7).**

2.6 GATA4 occupancy modulates FOXA2 high frequency binding spectrum minimally

Given the limited ability of the epigenome to explain specific FOXA2 binding choices, we speculated that high frequency binding may be directed by additional cofactors that have cell type specific expression. Cooperativity of TFs for target occupancy is common among non-pioneer TFs [5,155] and has recently been also demonstrated for pioneer factor occupancy [19]-[18]. To identify potential co-factors, we searched for differentially enriched motifs between genomic regions bound by FOXA2 exclusively in dEN or BJ^{FOXA2} (henceforth referred to as dEN exclusive or BJ^{FOXA2} exclusive) sites and cross-referenced this list against RNA-seq data for expression of their corresponding TFs (**Figure 2.10** and **Appendix S8a**). Motif sequences for several known endodermal regulators are enriched specifically at dEN exclusive sites, including motifs for regulators known to cooperate with FOXA factors. For instance, GATA4 binds to the ALB enhancer locus with FOXA2 in early gut endoderm cells prior to ALB expression [9,135,177]. Likewise, HNF4A, HNF1A and FOXA2 can convert human fibroblasts into hepatocyte-like cells [151,178].

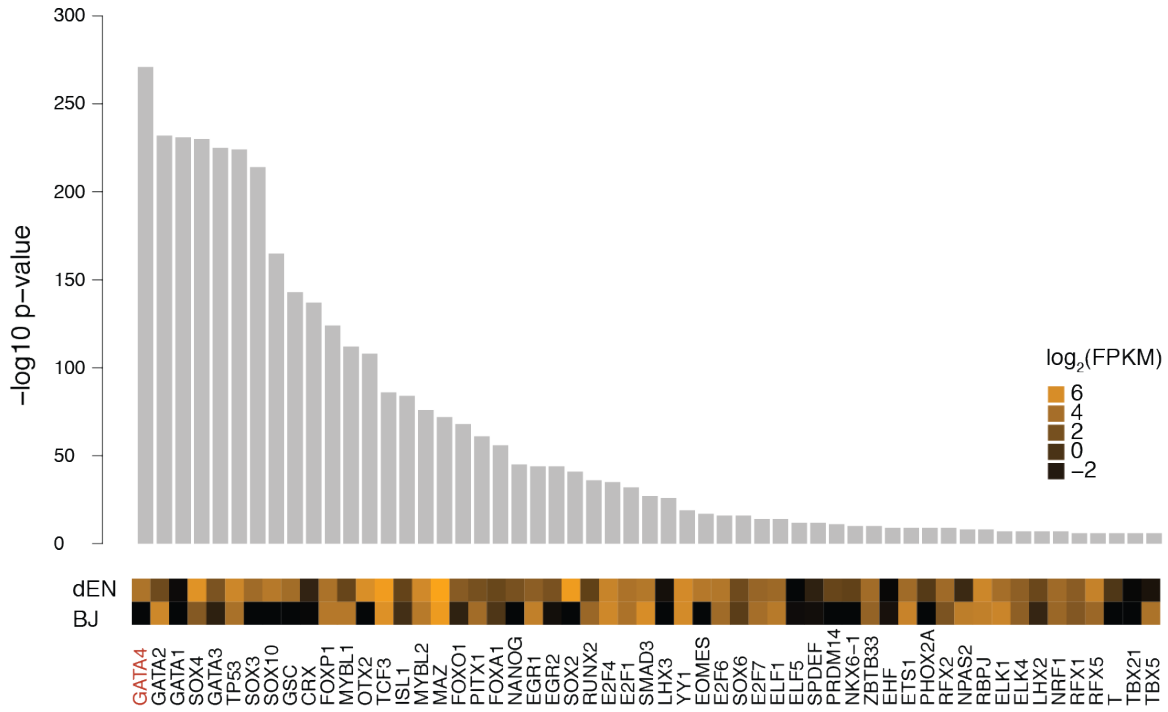


Figure 2.10:
Differential motif analysis displaying $-\log_{10}$ p-value of enriched motifs in dEN exclusive sites versus BJ^{FOXA2} exclusive sites with the most significant motifs on the left. Expression ($\log_2$ FPKM) of the TFs associated with the listed motif in both BJ^{FOXA2} and dEN. Of note, while there are many significant differential motifs observed in dEN exclusive sites, not all motifs are associated with factors that display differential expression.

Based on this analysis we selected GATA4 as a candidate co-factor that might influence FOXA2 binding behavior in the ectopic system based in its co-localized binding in liver and its ability to act as a pioneer factor itself [8,179]. We first used our previously published data [179] for FOXA2/GATA4 binding in dEN and found the two factors co-localized at 2,364 genomic sites, the majority of which overlap with dEN exclusive targets that are not highly enriched for FOXA2 in BJs (n=2,093 'dEN exclusive co-bound sites' **Appendix S8b**). To determine if GATA4 co-expression could modulate the high-enrichment binding spectrum of FOXA2 in the BJ^{FOXA}, we infected our BJ^{FOXA2}

line with a second lenti-viral construct containing constitutively expressed, V5-tagged GATA4 and induced FOXA2 with doxycycline for simultaneous expression of both factors for four days (BJ$^{FOXA2/GATA4}$; **Figure 2.11**).

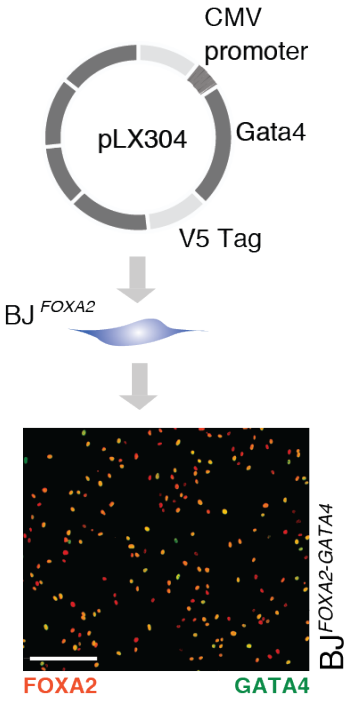
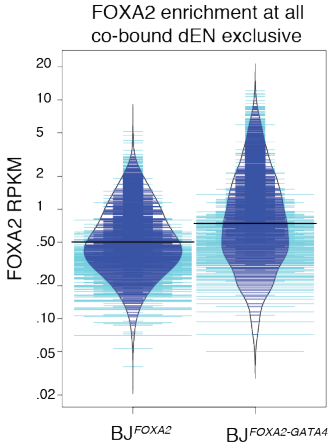


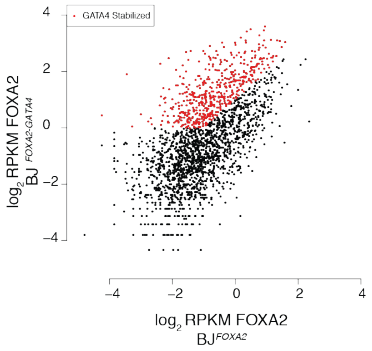
Figure 2.11:
Simplified schematic of ectopic expression system used to co-express GATA4 in BJ^{FOXA2}. Immunostaining of FOXA2 and V5-GATA4 in co-infected BJ^{FOXA2} fibroblasts. White scale bar is equal to 345nm.

We then performed ChIP-Seq for FOXA2 and found indeed an increased overall enrichment of FOXA2 and a significant FOXA2 enrichment at a subset of these targets (504 out of 2,093 – 'GATA4 stabilized' sites **Figure 2.12a**). Interestingly, GATA4 stabilized sites had slightly greater enrichment of FOXA2 enrichment prior to GATA4 expression signifying they were previously sampled by FOXA2 when expressed alone and indeed, we found the majority of these regions to be previously sampled by FOXA2 (63%; n=318 **Figure 2.12c-d; left bar plots**).

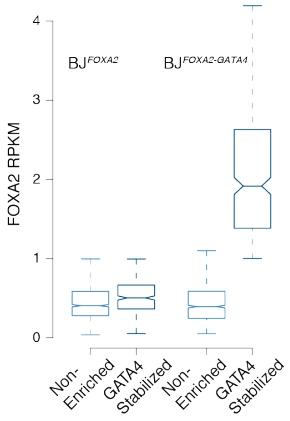
a.



b.



c.



d.

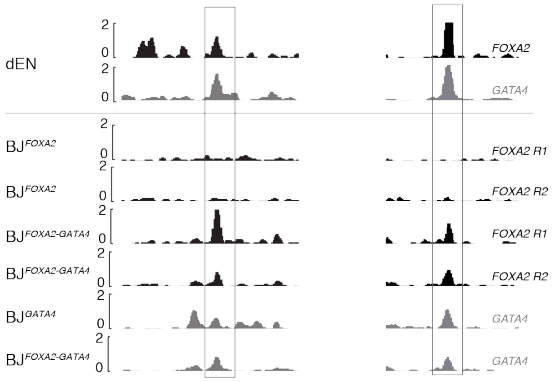


Figure 2.12:
a. Bean plot comparing FOXA2 enrichments at all dEN Exclusive co-bound sites when FOXA2 is expressed independently versus when FOXA2 and GATA4 are co-expressed. Thick black bars represent average. Blue lines indicate data points within the distribution while teal bars represent data points outside of the distribution.
b. Scatter plot comparing FOXA2 enrichment of co-bound dEN exclusive sites in BJ^{FOXA2} fibroblasts compared to BJ$^{FOXA2\text{-}GATA4}$ fibroblasts. Red dots indicate regions that gain at least 2 fold enrichment and are above RPKM of 1 in BJ$^{FOXA2\text{-}GATA4}$ fibroblasts.
c. Bar plots displaying the RPKM of FOXA2 enrichment in BJ^{FOXA2} and BJ$^{FOXA2\text{-}GATA4}$ at the subset of regions that are GATA4 stabilized compared to the non-enriched subset. Boxes indicate interquartile range and whiskers show maximum and minimum values. Outliers are removed.
d. Representative IGV browser tracks displaying FOXA2 and GATA4 enrichment in dEN and the various fibroblasts as indicated on the left. Gray bar highlights a region that shows the co-factor mediated recruitment of FOXA2 to two of its dEN targets chr13:76,031,782-76,039,815 and chr17:14,352,627-14,360,300

Also of note, we found most GATA4 stabilized sites were occupied by GATA4 when it is expressed independently (**Appendix S8c**). Given that GATA4 co-expression could only explain ~25% of the dEN exclusive co-bound FOXA2 targets, we searched for additional explanations and found the GATA motif differentially enriched in the GATA4 stabilized subset compared to the non-enriched subset (p-value 1.0e-5; motif occurring at 76% of regions). In turn, we observed the weak differential enrichment of other endodermal factor motifs in the non-enriched subset (T-box; p-value 1.0e-3, Eomes; p-value 1.0e-3; and SOX; p-value 1.0e-3) compared to GATA4 stabilized regions indicating stabilized occupancy of FOXA2 at these regions may be dependent on multiple TFs. Finally, similar to independent expression of FOXA2 and GATA4 alone we found that even in co-expression conditions, GATA4 stabilized targets showed little change in DNA accessibility compared to uninduced controls indicating that recruitment of chromatin remodeling machinery has not yet occurred and instead, these two factors co-localize on nucleosomes (**Figure 2.13).** In sum, these results support the model that the occupancy of the pioneer factor FOXA2 can partially be determined by cofactor engagement at specific subsets of genetically encoded target loci.

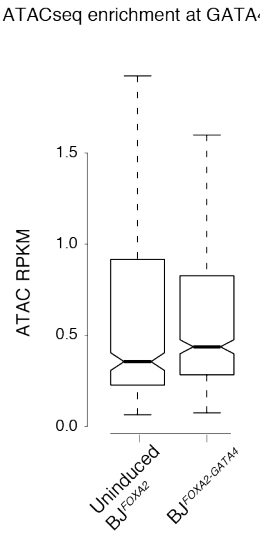


Figure 2.13:
Bar plots displaying RPKM of ATAC-seq enrichment in uninduced BJ^{FOXA2} versus BJ$^{FOXA2\text{-}GATA4}$ at GATA4 stabilized bites. Boxes indicate interquartile range and whiskers show maximum and minimum values. Outliers are removed.

2.7 Conclusions and Discussions

The work presented in this chapter represents the development of an engineered, ectopic system to examine genome-wide pioneer factor occupancy. We first compiled a set of *cis*-regulatory elements that are occupied by FOXA2 in endogenously expressing cell types and established that ectopic expression of FOXA2 at super-physiological levels in BJ fibroblasts cannot recapitulate the endogenous *cis*-regulatory spectrum observed across HepG2, A549 and dEN cells (**Figure 2.3)**. In fact, we observe two modes of genomic engagement by pioneer factors FOXA2 and GATA4. The first mode is characterized by cell-type specific, high frequency binding where we observe high enrichment of FOXA2 at distinct regions in BJ^{FOXA2}. The second mode is characterized by low-level sampling across the majority of regions occupied by FOXA2 in alternative lineages (**Figure 2.5)**. From these experiments we cannot distinguish if sampled sites are bound at high frequency in a small number of cells or if there is low

frequency binding at the same regions in every cell across the population as ChIP-seq signals are averaged across a population of cells.

Nevertheless, this characteristic appears to be unique to FOXA2 and GATA4 as in contrast, OCT4 occupancy displays high frequency binding alone. We do however, observe OCT4 sampling when OCT4 is co-expressed with its reprogramming factors SOX2, KLF4 and cMYC (**Appendix S4C**) which highlights distinct occupancy patterns of OCT4 when it is expressed independently versus in a reprogramming context. We propose that sampling of alternative target sites is a unique pioneer factor quality as we observe this for both FOXA2 and GATA4, which indicate that OCT4 may only act as a pioneer factor in specific cellular contexts, not universally.

Despite the sampling observed at FOXA2 targets in alternative lineages, we questioned why high frequency occupancy was reproducibly absent at these sites in our ectopic system. While pioneer factors by definition should not be restricted by the epigenetic landscape they encounter, OCT4 binding during the initial stages of reprogramming was observed to be restricted by heterochromatin domains [21]. Thus we set out to examine how the pre-existing epigenome influenced pioneer factor binding. We observed significant differences in the percentage of FOXA2 and GATA4 binding sites in pre-existing closed chromatin regions compared to OCT4 binding sites as the majority fell within pre-existing DNA accessible regions (**Figure 2.7**). This strengthens our suggestion that OCT4 alone may not retain the complete spectrum of pioneering capabilities that have been observed for it during reprogramming and thus we characterize OCT4 as a cooperative pioneer factor.

Because we observed minimal influence of the epigenome on FOXA2 and GATA4 occupancy, we next assessed if expression cell type specific TFs might modulate pioneer factor binding spectrum. We identified candidate TFs by first performing differential motif analysis on FOXA2 regions that are exclusively occupied in dEN (not occupied at a high frequency in BJ^{FOXA2}) and then by assessing the expression of factors associated with the differential motifs in dEN and BJ cells (**Figure 2.10**). We subsequently co-expressed, the dEN specific TF, GATA4 in BJ^{FOXA2} and re-examined FOXA2's occupancy in the presence of GATA4 expression. We find that a subset of previously categorized dEN exclusive regions are now occupied at high frequency in BJs suggesting that pioneer factor binding can be modulated through co-factor TF relationships (**Figure 2.11)**. Furthermore, the majority of these regions were previously sampled by FOXA2 in BJ^{FOXA2}. We find that GATA4 alone can occupy these regions, yet we observed little change in DNA accessibility upon both factors occupying these regions indicating that neither factor recruits chromatin-remodeling machinery and thus, we do not believe that this is dynamic assisted loading of FOXA2 at these target sites by GATA4. Instead, because both GATA and FOXA motifs are present at these sites, we suggest that GATA4 merely stabilizes enrichment of FOXA2 at these target sites in nucleosomal DNA. This could be via a direct or an indirect mechanism as our data cannot assess the differences. We reason that we do not observe stabilization at all dEN FOXA2 targets in BJs because multiple co-factors are likely to be involved in binding.

In conclusions, this chapter demonstrates that pioneer factors have a unique mode of genomic engagement where they display low-level enrichment at the majority

of their known target sites in alternative lineages. While OCT4 has previously been designated a pioneer factor, it may only have pioneer properties in certain contexts. Furthermore, because we utilized an ectopic system, we were able to assess and manipulate this system to understand more about pioneer factor occupancy. We focused on chromatin landscape and co-factor expression and find that we can minimally modulate FOXA2 occupancy by the expression of specific co-factors. While we are able to explain a portion of FOXA2 occupied sites, pioneer factor occupancy choice is convoluted and a large proportion of cell type specific sites are still not explained by our study. We did not examine all potential influences on occupancy such as chromatin confirmation, DNA shape features, and post-translational modifications to the TFs that may be playing a role in this process.

Chapter 3.
Determining the epigenetic and transcriptional impact of FOXA2 occupancy

Parts of this chapter are submitted for publication elsewhere [1].

3.1 Rationale

After DNA occupancy, the next critical piece in understanding how a repressed cis-regulatory element begins its activation process is to investigate the initial stages of epigenetic remodeling imposed directly by pioneer factor occupancy. Interestingly in Chapter 2, we found that the majority of regions targeted by ectopic FOXA2 were in regions of closed chromatin that contained DNAme and focused experiments and analysis on this subset of targets. While the epigenetic remodeling capabilities of FOXA factors have been explore in vitro [6,8] and during differentiation [24,25], this has limited genome-wide analyses to correlations between TF binding, nucleosome occupancy and histone modifications without distinguishing the molecular order critical to each discrete step. We focus on the initial gains in DNA accessibility as assessed by ATAC-seq along with the initial loss of DNAme observed at FOXA2 targets as assessed by ChIP-BS-seq. We hypothesized that utilizing a controlled ectopic system would enable us to move beyond correlative observations, to experimentally test mechanistic models of TF function in a targeted fashion allowing us to examine the effects of pioneer TF action on a temporal scale.

3.2 Global transcriptional and epigenetic impact of ectopic FOXA2 binding

To determine the molecular effects of the ectopic TF binding in the BJ^{FOXA2} cells we performed RNA-seq, ATAC-seq and ChIP-seq (H3K4me2 and H3K27ac) 48 hours post FOXA2 induction. In line with previous studies on pioneer factors [2,25], we find only a small number of genes that were immediately activated upon induction (~299 genes up-regulated and 191 genes down-regulated) and an even smaller number of the

activated genes appeared to have a FOXA2 binding site within 1kb of the associated gene promoter (~82 genes; **Appendix S9**). Due to the limited trans-activating properties observed for the factor, we instead focused on changes in chromatin upon FOXA2 occupancy. Globally, we observe regions that gain H3K4me2 and H3K27ac and further subset these regions into two sets: *de novo* regions that have minimal levels of either modification prior to FOXA2 occupancy, gain at least 2-fold signal as well as become enriched above RPKM = 1 and an *enhanced* set that already have enrichment for either mark and gain at least 2-fold more enrichment upon occupancy (**Figure 3.1**). Such gains in H3K4me histone modifications upon occupancy of FOXA factors have previously been observed as pioneer factors are known to establish competency at *cis*-regulatory regions [2,13,24,180]. Interesting, we also observed that ~40% of regions that gain H3K4me2 concomitantly gain low enrichment of H3K27ac (n= 1,937) indicating that FOXA2 can promote the establishment of active *cis*-regulatory elements.

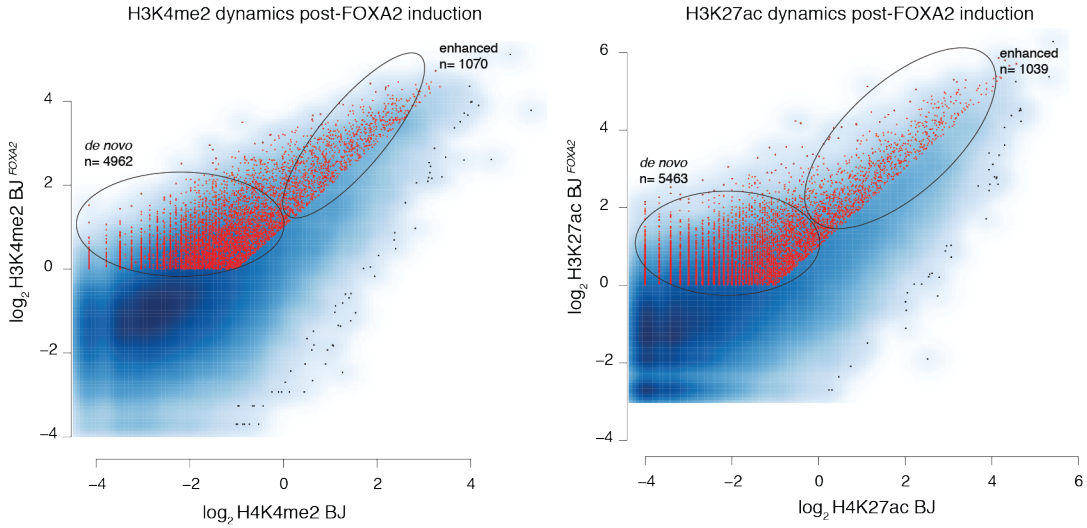


Figure 3.1:
Scatter plots of H3K4me2 and H3K27ac signal at pre- versus post- FOXA2 induction. Dots highlighted in red are at least 2-fold upregulated and become at least RPKM of 1 post- FOXA2 induction. Circles roughly highlight *de novo* gained versus enhanced changes.

3.3 DNA accessibility dynamics upon ectopic FOXA2 binding on repressed *cis*-regulatory elements

To understand a pioneer factors ability to induce overall remodeling at repressed *cis*-regulatory regions, we focused on FOXA2 occupied regions that fall in pre-existing closed chromatin (as defined by ATAC RPKM <1), assessed changes in DNA accessibility and correlated gains in active histone modifications. We first plotted the ATAC-seq signal (post FOXA2 induction) against the enrichment of FOXA2 at the entire FOXA2 union peak set (**Figure 3.2**).

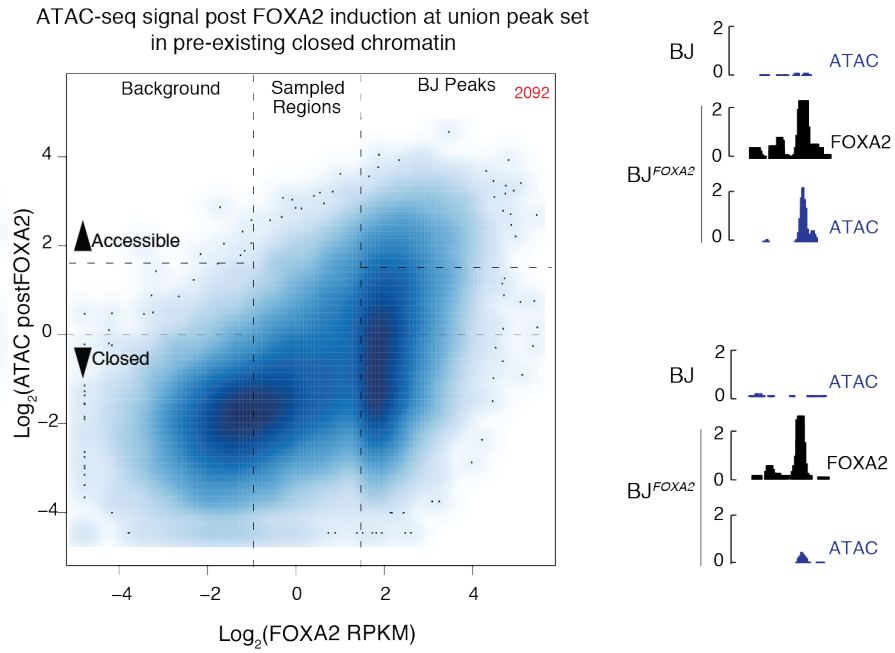


Figure 3.2:
Scatter plot displaying FOXA2 enrichment in induced BJ^{FOXA2} compared to post-FOXA2 induction ATAC-seq signal at the union set of FOXA2 binding sites that were considered to overlap closed chromatin (ATAC RPKM <1). Vertical lines separate the union set of FOXA2 peaks into background, sampled and called FOXA2 peaks. Horizontal lines indicate regions that have ATAC RPKM >3 that are considered accessible and ATAC RPKM <1 (considered closed). Representative IGV browser tracks of regions that become accessible upon FOXA2 binding (top; chr20:36,008,193-36,009,335) versus regions that remain closed (bottom; chr19:1,867,722-1,868,322).

First, as observed in the mid-range of the plot, the majority of the sampled regions show little change in ATAC signal suggesting that high enrichment binding is necessary for subsequent epigenetic remodeling. However, occupancy alone is not sufficient to induce significant changes in DNA accessibility as only a fraction (~13%) of high frequency FOXA2 regions (right side of plot) show significant gains in accessibility (n=2,092; **Figure 3.2**) while the majority of stably targeted regions showed minimal change in ATAC-seq signal post-induction. To determine whether any the observed regions that remain closed upon FOXA2 occupancy are in fact gene regulatory elements in any other cell type, we compiled a large number of the Roadmap Epigenomics Project DNase hypersensitivity data and found that 5,144 of 8,443 (requiring at least 20% overlap) of these sites do become accessible chromatin regions in at least one of cell types. It is also worth noting that we do observe low levels of increased ATAC-seq signal even at the target sites that remain closed based on our thresholds (**Figure 3.3a**). Indeed composite plots demonstrate a sharp, centralized gain in ATAC-seq signal at the peak center of the regions that become accessible and a smaller, but clearly visible, centralized gain at regions that remain inaccessible (**Figure 3.3b**). Thus FOXA2 has some measurable effect on most of its bound regions.

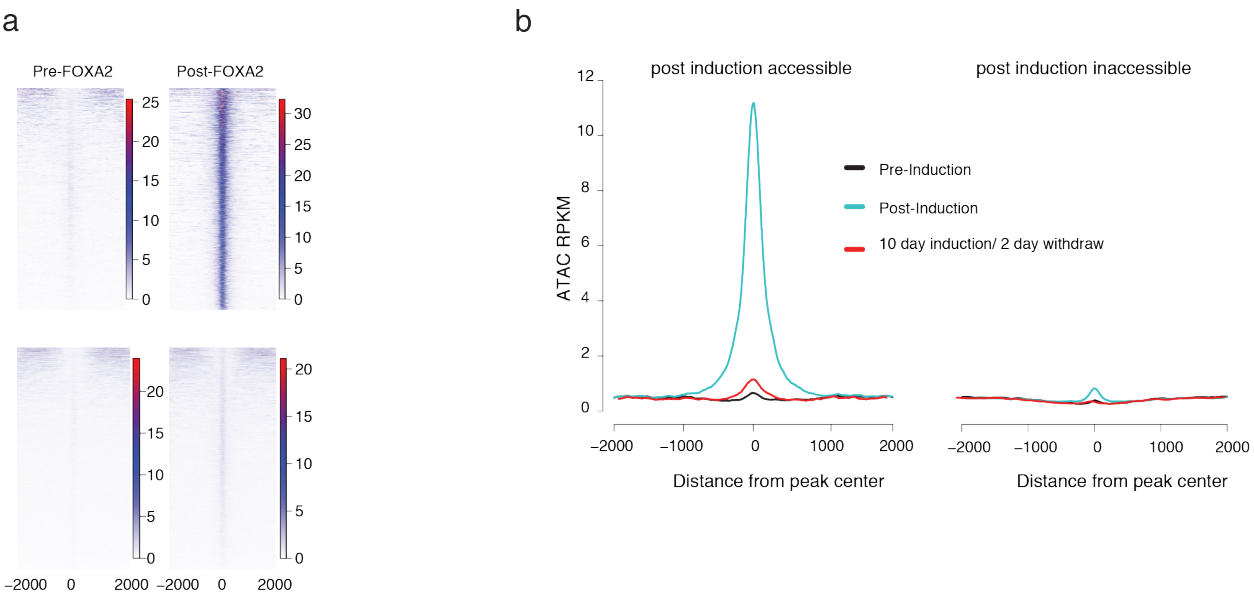


Figure 3.3:
a. Read density heat maps of post- FOXA2 ATAC-seq signal across all regions that become accessible (top) compared to the ones that remain closed upon FOXA2 binding (bottom).
b. Composite plots of ATAC-seq signal pre- and post- FOXA2 induction as well as after 10 days of DOX followed by 2 days withdrawal at regions that become accessible (left) and remain inaccessible (right) in BJ^{FOXA2}.

The FOX motif is more highly enriched in the subset of regions that become accessible (**Figure 3.4**, top 8 motifs shown) raising the possibility that they contain multiple FOX motifs, which in turn, may explain the minimal increased FOXA2 enrichment observed (**Figure 3.4**). In general, regions that become accessible have FOXA motifs more widely distributed across peak summit compared to regions that remain closed which display a more centralized motif occurrence (**Figure 3.4).** We also observed that the mean ATAC-seq signal before FOXA2 binding is higher in the sites that become accessible suggesting prior, yet minimal, accessibility in these regions (**Figure 3.4**).

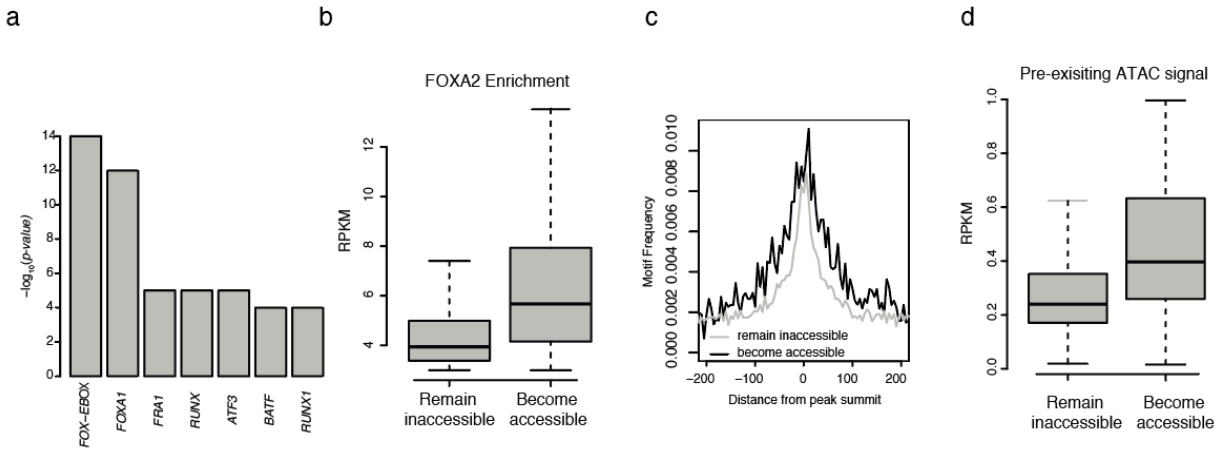


Figure 3.4:
a. Differential motif analysis as a bar plot using Homer for regions that become accessible versus regions that remain closed.
b. Mean enrichment of FOXA2 (RPKM) at regions that remain inaccessible versus those becoming fully accessible. Boxes indicate interquartile range and whiskers show maximum and minimum values. Outliers are removed.
c. Composite line plot of FOXA motif frequency across peak regions in those that become accessible (black) compared to the inaccessible set (grey)
d. Mean enrichment (RPKM) of pre-existing ATAC-seq signal at FOXA2 target site that remain closed and become open. Boxes indicate interquartile range and whiskers show maximum and minimum values. Outliers are removed.

Notably, by taking advantage of our doxycycline inducible system, we found that the majority of ATAC-seq signal gained after FOXA2 binding is already lost within 2 days of factor withdrawal indicating the transient behavior of this remodeling that has been previously observed (**Figure 3.3 and Appendix S10; red line)** [3]. Finally, we observe the occurrence of modified, phased nucleosomes surrounding the FOXA2 peak summit when we examine H3K4me2 and H3K27Ac in sites that become accessible compared to pre-induced BJ^{FOXA2} (**Figure 3.5).** Regions that remain inaccessible do not demonstrate significant histone enrichment compared to pre-induced BJ^{FOXA2} indicating

that accessibility correlates with deposition of active histone modifications post-FOXA2 occupancy (**Figure 3.5**).

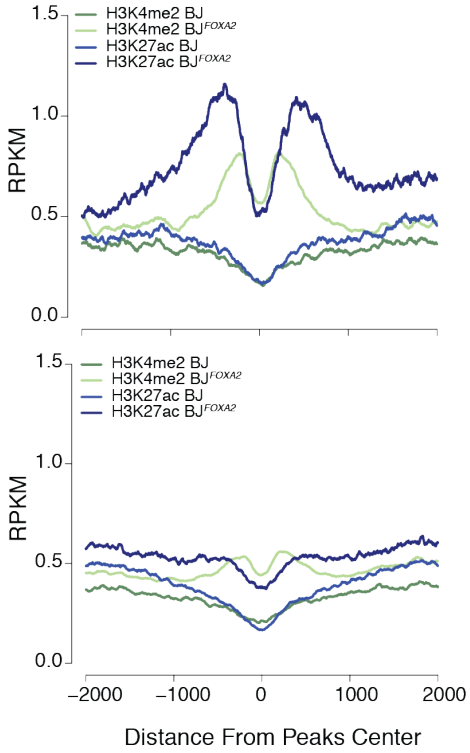


Figure 3.5:
Composite plots displaying H3K27ac and H3K4me2 enrichment over regions that become accessible and those that remain closed comparing pre- and post- FOXA2 occupancy.

In fact, we ranked all BJ peaks found in pre-existing closed chromatin (ATAC RPKM <1) by post FOXA2 ATAC-seq enrichment signal, evenly binned regions and calculated the ATAC-seq mean for the bin and the corresponding mean for H3K4me2 and H3K27ac signal post-FOXA2 induction. This demonstrates a somewhat linear relationship between gain in DNA accessibility and gain in H3K4me2 (along with a weaker correlation for H3K27ac) post FOXA2 occupancy (**Figure 3.6).** It is worth stating that the enrichment of active histone modifications is modest and does not reach a similar enrichment level of active promoters for example.

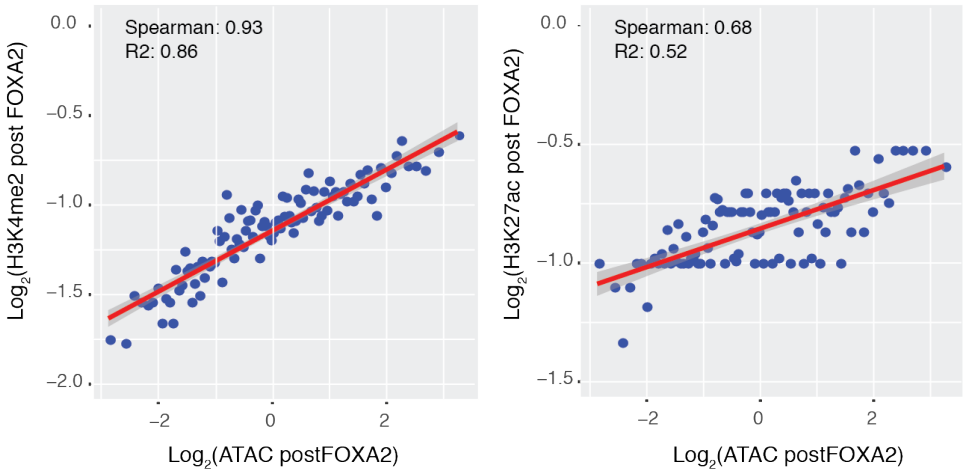


Figure 3.6:
Binned scatter plot for BJ peaks is pre-existing closed chromatin (ATAC RPKM > 1) comparing ATAC-seq and H3K4me2 or H3K27ac signal post- FOXA2 induction. Note that mean signals are calculated for the 600bp window that surrounds the FOXA2 peak summit and this windowing size only partially captures modified histone enrichment.

3.4. FOXA2 targets display unique DNA methylation dynamics

The majority of FOXA2 occupied regions fall in areas containing DNAme (**Figure 2.7)** and previous studies found that FOXA2 binding is strongly associated with loss of cytosine methylation [24,25]. Given the localized and distinct DNA methylation dynamics observed across enhancers during differentiation and development [25,93], we chose to further explore the loss of DNAme at FOXA2 targets using our ectopic BJ^{FOXA2} system and thus performed ChIP followed by bisulfite sequencing (ChIP-BS-seq) to quantify DNA methylation levels on fragments that were physically associated with FOXA2 [181]. For the analysis we included methylation levels of CpG dinucleotides captured at ≥3X in the BJ^{FOXA2} ChIP-BS sample and compared them with matched CpGs from BJ^{FOXA2} WGBS data prior to FOXA2 induction. We found that FOXA2 occupies three distinct sets of genomic regions: those in pre-existing hypomethylated DNA (binding site mean

methylation < 20%) that, not surprisingly, remained hypomethylated after FOXA2 binding (**Figure 3.7 and 3.8**, Class 1: n=16,742), those that display high methylation levels before and after FOXA2 binding (**Figure 3.7 and 3.8** Class2: n=8,794) and a unique class of regions that display a clear loss of DNAme (at least 20%, and often more) change in mean methylation following the binding of FOXA2 (**Figure 3.7 and 3.8** Class 3: n=9,111).

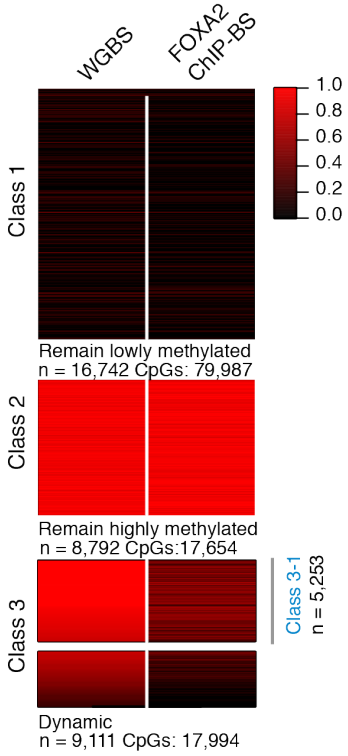


Figure 3.7:
Heat map of CpG methylation levels for matched CpGs comparing BJ WGBS (pre-FOXA2 induction) and post-FOXA2 induction ChIP bisulfite sequencing data. Three main classes of FOXA2 binding emerge: Class 1 – remains lowly methylated (regions n=16,742), Class 2 – remains highly methylated (regions n=8,792) Class 3 – Dynamic (regions n=9,111). Gray bar in Class 3 indicates the subset of dynamic regions that have a pre-induction methylation level of greater than 80% (regions n=5253) and are further referred to as Class 3-1.

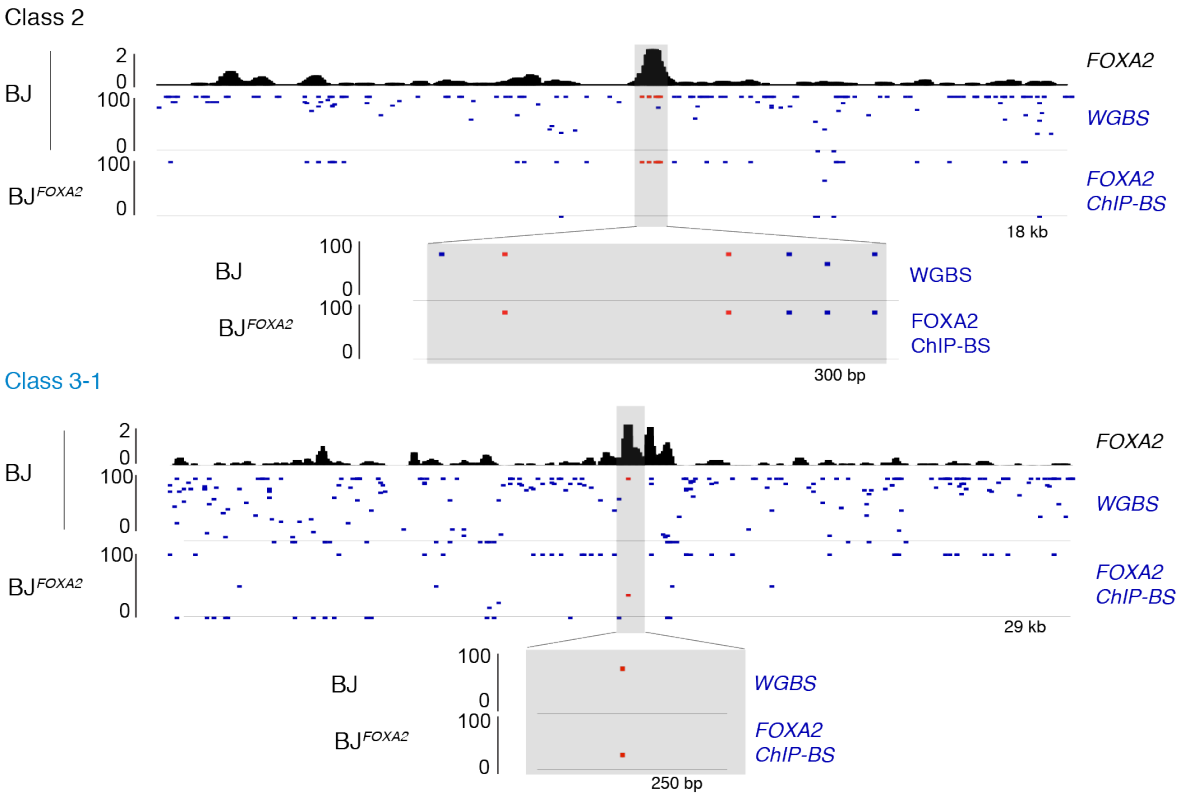


Figure 3.8:
Representative IGV browser tracks showing BJ^{FOXA2} FOXA2 enrichment, CpG methylation pre- FOXA2 induction and FOXA2 ChIP-BS data post induction. Top half is an example of a class 2 regions (hypermethylated) located at chr12:54,002,592-54,021,127. The bottom shot is an example of a class 3 region (dynamic) located at chr18:32,911,411-32,941,267. Violin plot of CpG density of class 2 and class 3 target sites. CpG density is calculated by the number of CpG dinucleotides across 100bp windows divided by total number of base pairs. Dots represent median values while black lines indicate interquartile range.

As most studies have observed FOXA targeted regions overlapping with areas of low DNAme [24,25], it was a bit unexpected to observe FOXA2 occupying regions that retain high methylation post binding, which prompted us to scrutinize the differences between class 2 and 3 target sites more closely. First, we used *in vitro* electro-mobility shift assays to demonstrate that FOXA2 directly interacts with methylated, hemi-methylated and unmethylated DNA species (**Appendix S11).** We then selected a more stringent subset of class 3 targets with initial methylation of at least 80% in uninduced

BJ^{FOXA2} (**Figure 3.7**; Class 3-1, n = 5,253) to be more comparable to our class 2 subset of targets that contain mean methylation values above 80%. First, we do not observe any correlation between change in DNAme and FOXA2 enrichment at class 3-1 targets indicating that increased FOXA2 binding is not responsible for greater loss of DNAme (**Appendix S12**). Both groups are indistinguishable in terms of overlap with common genomic features and CpG density (**Figure 3.9a** mean CpG count 4.2 and 4.8 for class 2 and 3-1 respectively). However, upon closer inspection we found that the class 2 target sites that remain hypermethylated were comparatively more depleted of CpG dinucleotides near the peak summit given the average sequencing coverage across these two groups was indistinguishable (**Figure 3.9b**).

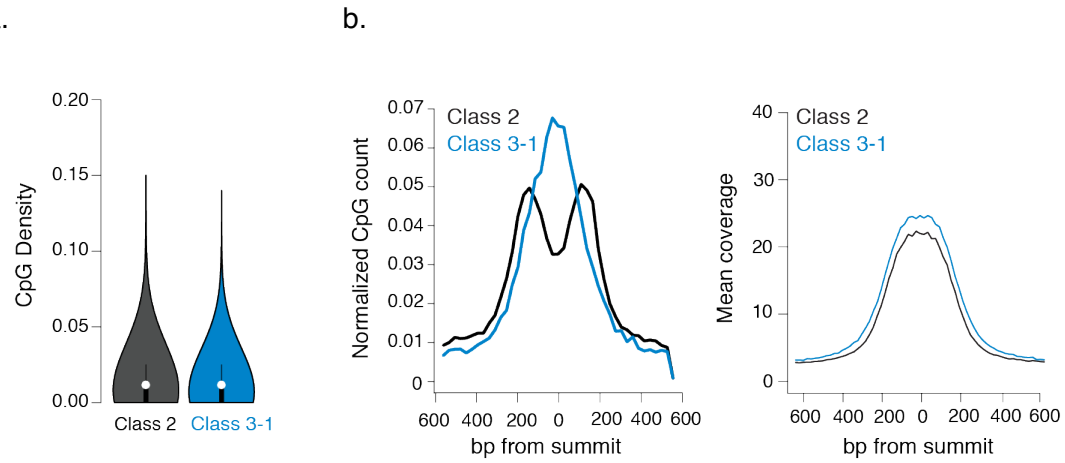


Figure 3.9:
a. CpG density for class 2 and class 3-1 targets
b. Composite plots showing normalized CpG count of Class 2 and Class 3 target sites (left). Class 2 is depleted for CpGs toward the center of the peak while Class 3 targets are enriched. Mean sequencing coverage between Class 2 and Class 3 target sites is equivalent (right).

In line with this observation, we calculated the distance from the summit to the nearest CpG and found it was significantly larger for class 2 targets versus class 3-1 targets (**Appendix S13**; average 73.8bp and 90.2bp, respectively $p < 2.2^{-16}$), while the

average methylation levels in uninduced BJ^{FOXA2} showed no significant difference (**Appendix S13;** p = 0.95, average 95% methylated in both). Furthermore, CpGs across the whole binding site were increasingly demethylated towards the summit center compared to pre FOXA2 methylation levels (**Figure 3.10** and **Appendix S14**), confirming the dependence of dynamic loss of methylation on distance to the center of the FOXA2 binding site.

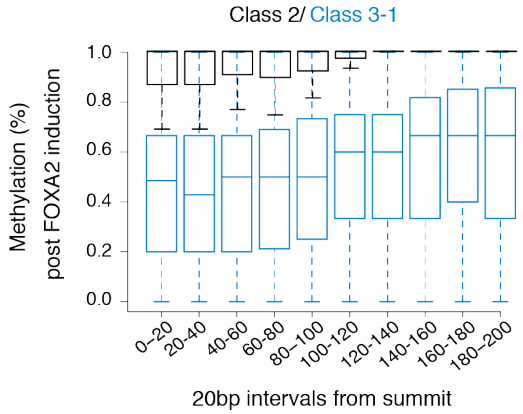


Figure 3.10:
Box plot show the percent methylation of CpGs within 20bp windows from the summit of the peak and extended up to 200bp. Methylation measurements were taken from ChIP-BS data after FOXA2 induction. Class 2 black. Class 3-1 blue. Boxes indicate interquartile range and whiskers show maximum and minimum values. Outliers are removed.

When we reduced the window size surrounding the FOXA2 peak summit from 200bp down to 20bps we still found a large percentage of bound regions are in areas of hypermethylated DNA (20.9% with 200bp window versus 13.8% with 20bp window). Based on these results we cannot distinguish whether class 2 target sites are merely a consequence of the imprecise peak calling that would disappear with a more refined mapping experiment such as ChIP-Exonuclease [182] or whether there is indeed a

functional difference between class 2 and 3 regions. We found that class 2 regions are less likely to gain ATAC-seq signal post-induction compared to class 3-1 although even target sites that remain methylated (class 2) can gain some level of ATAC signal post FOXA2 occupancy (**Appendix S15**). While we observe a minimal gain in H3K4me2/H3K27ac signal post FOXA2 induction, we observe no difference in this enrichment between class 2 or 3 target sites, and in fact, neither class shows significant enrichment post FOXA2 induction (with mean values around .5 RPKM; **Appendix S16**), suggesting that loss of DNAme is unlikely to occur because of the presence of recruited histone modifying enzymes and may be due to FOXA2 occupancy directly. In fact, we individually deleted both the n- and c-terminal domains of FOXA2, performed ChIP-BS-seq for these constructs and find that class 3 target sites can still dynamically lose DNAme signifying that neither the n- nor c- terminal domain is responsible for the recruitment or interaction with a demethylating agent (**Appendix S17).** Due to the striking dynamic change in DNA methylation at class 3-1 targets, we focused our attention on the underlying mechanisms of FOXA2 induced loss of methylation (Chapter 4).

3.5. Discussion and conclusions

Global observations at FOXA2 binding sites revealed many more changes to the chromatin than global transcriptional changes. Overall, we could only associate 82 FOXA2 binding instances with up-regulated transcription even though we observed close to 300 up-regulated genes. We were only able to associate a FOXA2 binding event with a gene if it was within 1kB distance of the promoter of a gene. Though we are not completely sure that this regulatory element actually has an effect on the closest

gene, this is the standard practice in the field. The association of FOXA2 binding sites with a gene could be improved with chromatin confirmation data. This way, we would be able to associate distal regulatory elements with the promoter regions of genes to which they are looped in three dimensions. With this, we would likely be able to associate all 300 up-regulated genes, with one of more FOXA2 occupied distal regulatory elements and have a better appreciation of the global enhancer landscape. Furthermore, we also identified a smaller number of genes that are down-regulated upon FOXA2 induction and are not surprised by this given that FOXA has been known to associate with repressive complexes in development [183]. Nevertheless, we still identified thousands of regulatory regions that gain enrichment for H3K4me2 and H3K27ac and considering the major pioneer factor role in establishing competence at *cis*-regulatory elements we expected to observe greater changes to chromatin than transcription [2].

Our investigation into the consequence of FOXA2 occupancy specifically at repressed *cis*-regulatory elements revealed that surprisingly, only a small percentage of target sites display significant change in DNA accessibility. These sites contain greater enrichment of FOXA2, elevated differential FOXA motif occurrence and concomitantly gain significant enrichment of modified histones. This can be interpreted in line with previous *in vitro* data that demonstrated position of motif occurrence was important for nucleosome eviction [160]. Yet, almost all targets become, to some extent, more accessible indicating that the presence of FOXA2 binding must alter the association of the DNA with the nucleosome none-the-less. Given recent findings that active enhancer regions may still contain accessible nucleosomes within ATAC-seq positive regions

[3,90,91], it is possibly that some of the most significant changes in DNA accessibility are at regions that gain modified histones due to the low DNA stability of H3K4me2 and especially H3K27ac [90]. Though a chicken and the egg scenario arises here as we cannot distinguish whether the gain in active histone modifications is a result of significant gains in DNA accessibility or if their enrichment causes significant gains in accessibility. While we were unable to associate any pre-existing chromatin state feature to these regions that could explain their new accessibility, we admit that we have not mapped or observed histone variants such as H2A.Z or H3.3 in the BJ fibroblast system. H2A.Z was found co-localized at regions that become bound by FOXA in mouse endoderm development and gain DNA accessibility [162]. Furthermore, we did not measure the localization of linker histone H1 in the pre-existing chromatin state. FOXA2 was recently shown *in vivo* to outcompete linker histone H1 at active, liver specific enhancer regions resulting in an overall gain in accessibility [3].

We observed two distinct responses upon FOXA2 occupancy at highly methylated regions – either the regions remain highly methylated (class 2) or the regions dynamically lose DNAme (class 3). It appears that the distance of the FOXA2 binding site to the nearest CpG drives the distinction between these two different responses as we could not attribute any other pre-existing epigenetic, or sequence feature at these two classes of regions. After FOXA2 induction, we find that class 3 dynamic regions gain more DNA accessibility overall than class 2 regions (yet, we notice a modest gain in enrichment at class 2 regions as well). Importantly, we observe minimal yet equivalent gain in active histone modifications at class 2 and class 3 regions, which indicates that loss of DNAme is unlikely to occur because of direct

recruitment of histone modifying enzymes to these regions. This suggested that FOXA2 itself may cause the loss of DNAme at this subset of target sites and investigated the mechanism of target site demethylation in Chapter 4.

Chapter 4.
Mechanistic dissection of epigenetic remodeling imposed by FOXA2 occupancy

Parts of this chapter are submitted for publication elsewhere [1].

4.1 Rationale

A great deal of in vitro work has described the potent mechanism by which pioneer factors, specifically the FOXA family, act to remodel nucleosomes to access target sites in closed chromatin [2,6,8,9,25,51,162,177]. However, the regulatory underpinnings that direct DNA demethylation at silent loci is another critical component of gene activation that has been difficult to decouple from mapping TF binding – DNAme relationships within endogenous cell types. In Chapter 3 we demonstrated by ChIP-BS-seq that a subset of regions occupied by FOXA2 dynamically lose DNAme indicating a direct role for FOXA2 in the demethylation process. This chapter focuses on utilizing our ectopic system to decipher the mechanism by which FOXA2 induces demethylation at its occupied regions. Two potential methods for demethylation are plausible in this context: a passive mechanism in which the depletion of 5-methyl-cytosine occurs following subsequent DNA replication or an active enzymatic removal of 5-methyl-cytosine.

4.2 Ectopic system halting DNA replication

Transitioning a cytosine from the methylated to the unmethylated state requires either an active, enzymatic removal of the methyl group [115], a passive, replication dependent loss which would require blocking the maintenance methylation activity of DNMT1 at the specific cytosines following synthesis of the nascent DNA or a combination of both [92,109]. To investigate and attempt to distinguish between these possibilities, we designed an experimental strategy in which we chemically halted DNA replication by blocking the BJ^{FOXA2} cells in G1 with mimosine treatment. We then

released half of the cells back into normal replicating conditions by washing out the chemical treatment while the other half persisted with mimosine treatment to maintain the G1 block. We simultaneously induced FOXA2 for 24h in both samples and again, used EdU incorporation to verify the cell cycle effects and performed FOXA2 ChIP-BS-sequencing (**Figure 4.1**).

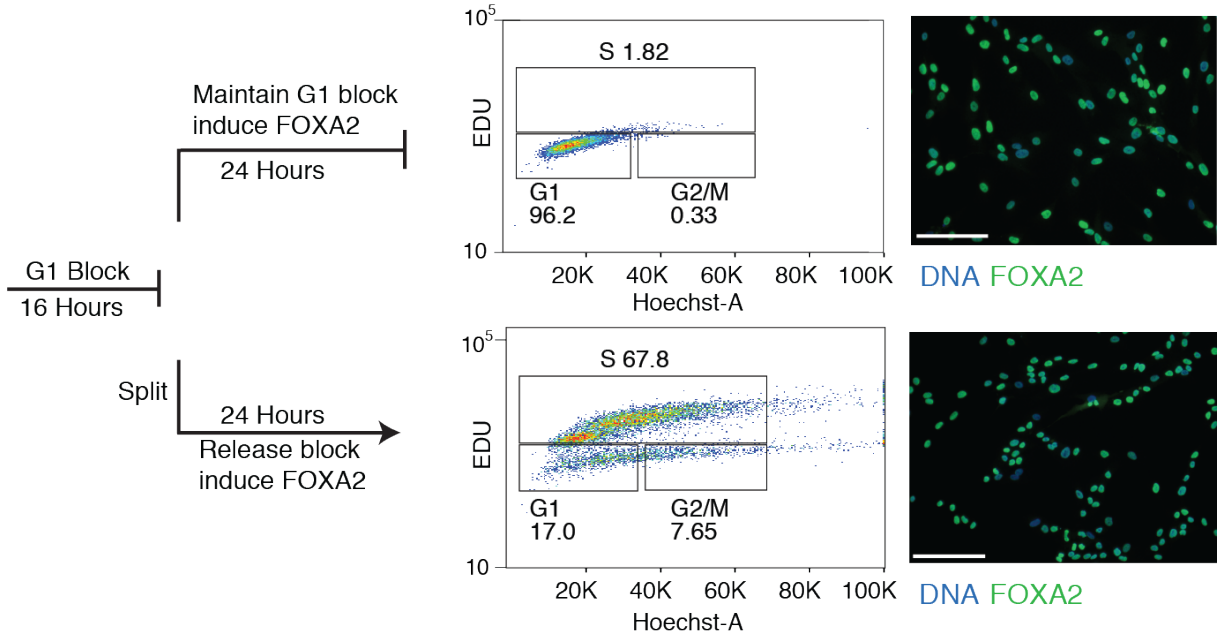


Figure 4.1:
Design of the DNA replication experiment to obtain two different populations for FOXA2 ChIP-BS analysis. FACS analysis using EdU incorporation and Hoechst DNA stain shows that halted cells remain in G1 and cells that are removed from chemical block can proliferate at a normal rate. Both populations of cells express FOXA2 highly as shown by immunostaining on the right. White scale bar is equal to 345nm.

During that time window the non-arrested cells have approximately gone through 1-2 rounds of cell division based on carboxyfluorescein succinimidyl ester (CSFE) labeling (**Appendix S18**). Notably, FOXA2 targeted similar genomic regions in both the cell

cycle halted and normal replicating conditions indicating that even in non-replicating cells, FOXA2 protein can accumulate (**Figure 4.1)** and access similar DNA targets (**Figure 4.2**).

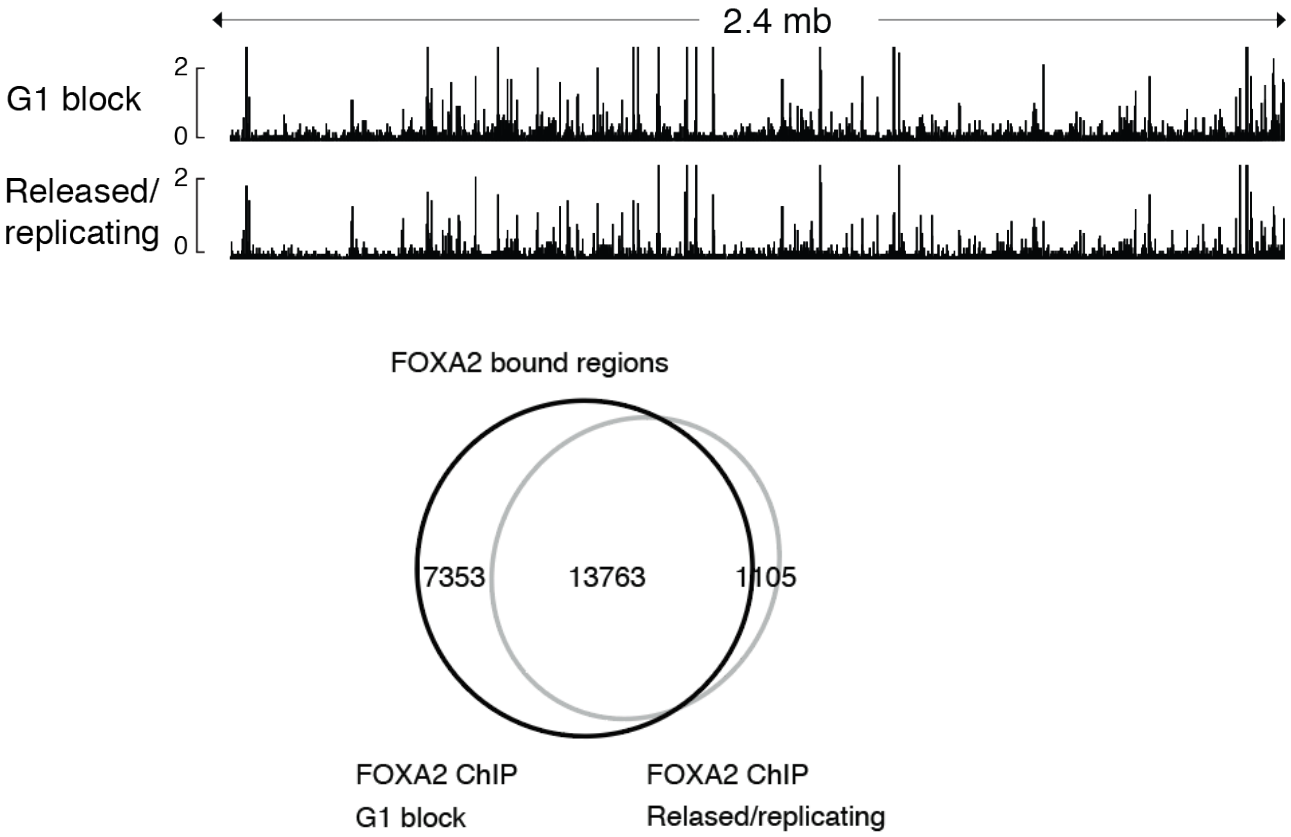


Figure 4.2:
Representative IGV browser tracks of a 2.4mb window (chr8:57,156,834-59,597,409) showing FOXA2 ChIP-seq data from cells halted in G1 (top track) and cells that were halted and then released back in to normal cycling conditions (bottom track). FOXA2 binding and accumulation is visually similar in both experiments. IDR peak calls were made for these two FOXA2 ChIP-sequencing experiments.
Venn diagram of the overlap reveals high similarity in called peaks between the two samples.

4.3 Loss of DNA methylation but not occupancy nor nucleosome remodeling is dependent on DNA replication

Quite strikingly however, the dynamic loss of DNAme at the FOXA2 occupied regions was only observed under the replicating conditions (**Figure 4.3, left**). Cells that remained halted in the G1 block, no longer displayed any dynamic change in DNAme levels and instead, remained highly methylated suggesting the observed loss of DNAme is, at least in part, dependent on DNA replication. In contrast, we find that FOXA2's ability to alter DNA accessibility upon occupancy is not dependent on DNA replication as the regions that we previously observed as gaining significant accessibility (**Figure 3.2;** n= 2,092**)** have highly correlated ATAC-seq enrichment in G1 blocked cells versus replicating cells (**Figure 4.3**; **right** pearson 0.84).

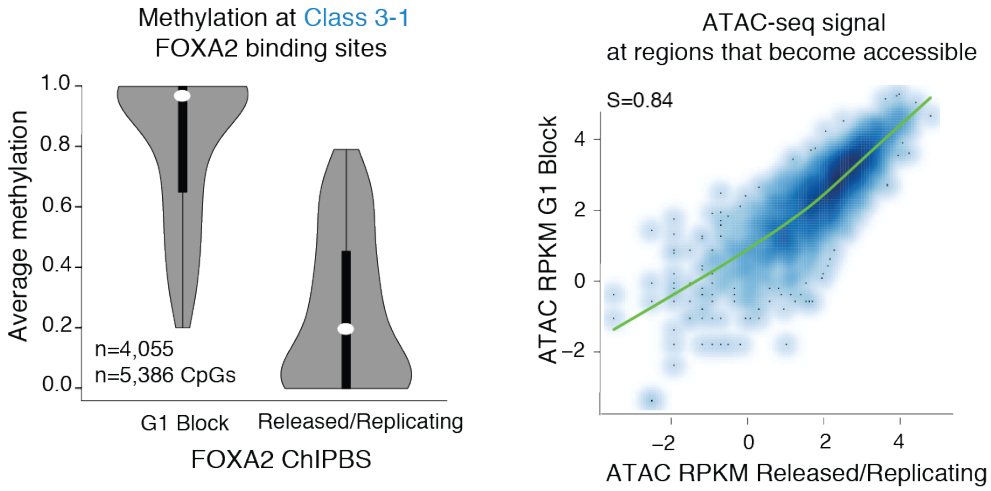


Figure 4.3:
Violin plots of the average methylation levels for Class 3, 'Dynamic' peaks between cells halted in G1 and cells in normal replicating conditions. Dynamic regions no longer lose DNAme when cells are halted in G1. Dots represent median values and bar indicates interquartile range.
Scatter plot of ATAC-seq signal post FOXA2 induction at regions that become accessible upon FOXA2 binding in G1 halted cells compared to cells released back into normal cycling conditions. Spearman correlations = .84

4.4 FOXA2 depletion in S-phase disrupts dynamic DNAme loss

Therefore, we observed that FOXA2 can physically associate with the DNA in both non-replicating and replicating conditions (**Figure 4.2)**, yet loss of DNAme is only observed after the cells undergo at least one round of DNA replication. Because DNAme patterns have to be copied onto nascent strands after DNA replication, we hypothesized that the immediate recruitment of FOXA2 to DNA target regions following DNA replication (S-phase) may be sufficient to block maintenance methylation of DNMT1. To assess this mechanism, we generated BJ fibroblasts expressing FOXA2 fused to CDT1 (BJ$^{FOXA2\text{-}CDT1}$) to specifically deplete FOXA2 expression during S-phase of the cell cycle **(Figure 4.4**, ref [184]). Western blot quantification of FOXA2 protein levels during the cell cycle indicates differential expression in G1 compared to S/G2/M. However, we observe moderate protein levels during S/G2/M that may, in part, be the result of the super-physiological expression (**Figure 4.4**).

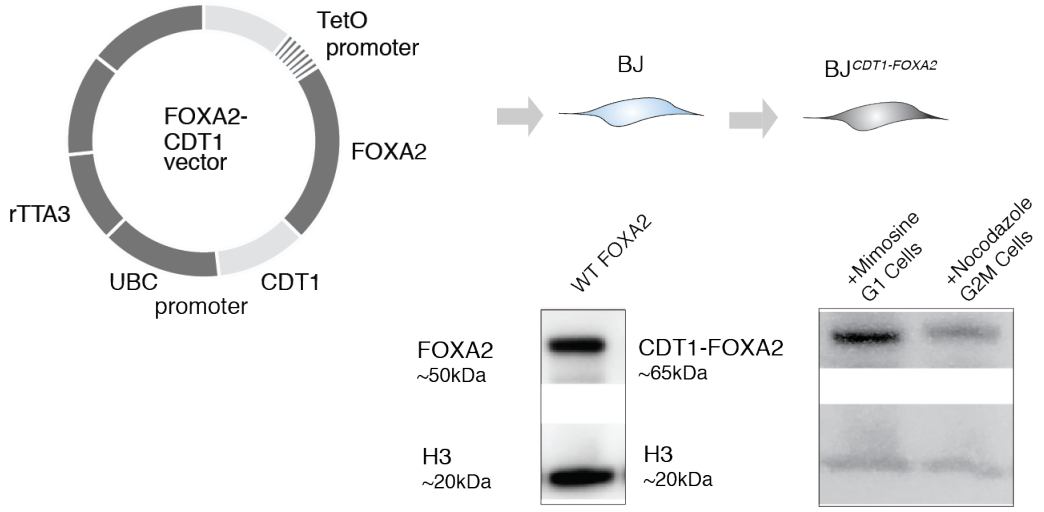


Figure 4.4:
Schematic representation of FOXA2-CDT1 fusion lentiviral construct generated with corresponding western blot for FOXA2 in BJ$^{FOXA2\text{-}CDT1}$ cells treated with Mimosine to enhance proportion of cells in G1 and Nocadazole to enhance proportion of cells in S/G2/M. H3 levels shown as loading control. And wildtype protein levels are also shown as a control.

Nevertheless, we then induced FOXA2 for 4 days, performed ChIP-BS-seq for FOXA2, and first observed an overall high correlation of FOXA2 enrichment in BJ$^{FOXA2\text{-}CDT1}$ compared to BJ^{FOXA2} (**Appendix S19**). We next examined levels of DNAme at previously defined dynamic regions that are highly covered in both samples and observe a decreased loss of DNAme in BJ$^{FOXA2\text{-}CDT1}$ cells compared to wildtype BJ^{FOXA2} ChIP-BS results at both class 3 and 3-1 regions (**Figure 4.5 Appendix S20).** Taken together, our data indicate that S-phase binding of FOXA2 may be required for the loss of DNAme observed at our dynamic FOXA2 target regions. While we still observe some loss in DNAme at class 3 regions, we suspect that this could be due to residual FOXA2 protein levels during S/G2/M (**Figure 4.4)**.

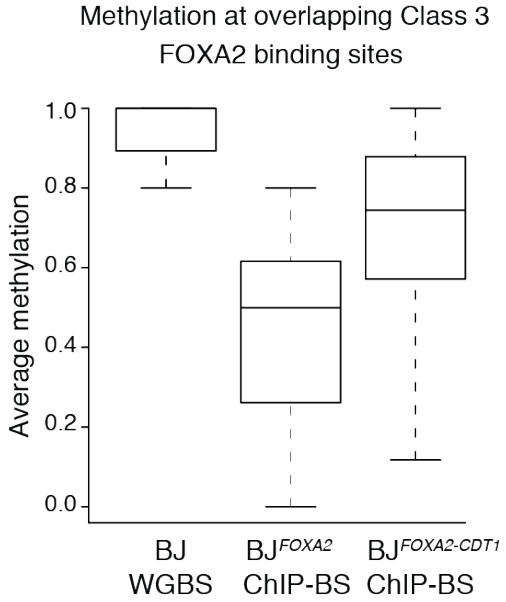


Figure 4.5:
Box plots show average methylation of all dynamic regions with initial hypermethylation (Class 3-1) in BJ WGBS, BJ^{FOXA2} ChIP-BS and BJ$^{FOXA2\text{-}CDT1}$ ChIP-BS data. Regions shown had at least 10X coverage. Boxes indicate interquartile range and whiskers show maximum and minimum values. Outliers are removed.

4.5 Discussion and conclusions

Mechanistic investigations into the loss of methylation seen at some FOXA2 occupied regions uncovered a dependence on DNA replication as cells halted in G1 fail to dynamically lose DNAme compared with wild type replicating cells. This suggested more of a passive loss of DNAme than an active, enzymatic pathway. Though we can still not rule out the possibility that this loss occurs by active demethylation that is also dependent on DNA replication. For instance, the active, oxidation of the cytosine base to 5hmC, could subsequently be depleted after multiple rounds of replications. The utilization of base excision repair to impose active demethylation (Chapter 1) would likely function during all phases of the cell cycle and not be inhibited with the blocking of DNA replication. A recent study speculated that loss of DNAme at FOXA1 targets resulted from an active demethylation mechanism by eventually involving DNA repair enzymes [185], but they did not examine DNAme loss in the absence of DNA replication nor report the active demethylating enzyme leaving several alternative possibilities open. Notably, we still detect strong FOXA2 occupancy at the identical target spectrum even when DNA replication is chemically halted in G1. This suggested that FOXA2 is unlikely to recruit anything to catalyze the loss of DNAme at these targets.

Thus we hypothesized that the binding of FOXA2 during S-phase of the cell cycle may induce the loss in DNAme. We subsequently examined DNAme upon FOXA2 occupancy in a FOXA2-CDT1 fusion cell line that displayed minimal protein levels of FOXA2 during S/G2/M and observed a decreased loss in DNAme levels at dynamic target regions. Recent studies of TF binding and DNAme suggested that NRF1 was unable to outcompete maintenance DNAme resulting in remethylation of target sites and loss of NRF1 occupancy [99]. In contrast, our study suggests that S-phase binding of

FOXA2 occurs rapidly after nascent strand synthesis prior to maintenance methylation, and this may be achievable due to an overall lag in DNA nascent strand remethylation observed after replication (unpublished data Meissner Lab).

In contrast, significant gains in DNA accessibility are observed regardless of active DNA replication. We are developing a potential model in which FOXA2 can occupy and immediately increase accessibility at target loci regardless of the cell cycle phase, yet its binding in S-phase of the cell cycle, soon after the passage of the replication fork, may block maintenance methylation resulting in the loss of DNAme at target sites. FOXA2 is likely to only block maintenance methylation at CpGs within or very close to its binding site highlighting the distinct in class 2 and class 3 target sites. However, more experiments are needed to demonstrate FOXA2 occupancy on nascent DNA strands to validate this model.

Chapter 5.
Discussion, Conclusions and Future Directions

Parts of this chapter are submitted for publication elsewhere [1].

5.1 Summary

Transcription factor (TF)-coordinated activation of repressed *cis*-regulatory elements is a key step towards gene activation and ultimately cell state transitions in development and reprogramming. Because pioneer TFs are distinguishable by their ability to occupy target sites in closed chromatin, they are considered on the top of the TF hierarchy as the initiators of gene activation. Despite these unique occupancy properties, accumulating data has suggested that *in vivo,* pioneer factors may behave similarly to non-pioneer factors by occupying cell-type specific genomic regions [16] and by demonstrating occupancy restrictions due to chromatin [21]. Thus we set out to test the abilities and limitations of pioneer TFs in an ectopic system that removes the factor from its normal developmental context (**Figure 2.2)**. We first created lists of *cis*-regulatory elements that are endogenously occupied by a set of previously defined pioneer factors, FOXA2, GATA4 and OCT4 across four diverse cell types so that we could determine if these regions would be equally occupied under super-physiological expression conditions in an ectopic cell type. We found that despite high protein levels of the pioneer factors in our ectopic systems, each factor occupied a mostly unique set of genomic regions with only a small percentage of ectopic and endogenous peaks overlapping (**Figure 2.3)**. Interestingly however, we also found that pioneer factors, FOXA2 and GATA4 displayed significant low enrichment indicating sampling by TFs at the majority of those sites that are uniquely enriched in alternative lineages, but these low enrichments did not reach the threshold at which they would be considered a significant increase compared to background (or a “peak”) in the ectopic system (**Figure 2.4 and 2.5).** In contrast, ectopically expressed OCT4 did not display a sampling

phenotype at regions uniquely occupied in alternative cell types and instead exhibited a distinct, bimodal distribution of TF enrichments compared to GATA4 and FOXA2 (**Figure 2.5).**

Next, to elucidate factors contributing to the unique pioneer factor occupancy spectrum observed across cell types, we assessed how the pre-existing epigenetic landscapes of the starting cell type and/or the expression of unique co-factors within the cell type might affect occupancy. We found that FOXA2 and GATA4 were able to bind to the majority of our pre-assigned chromatin states, whereas OCT4 displayed a clear bias for pre-existing open chromatin features (**Figure 2.7).** The majority of FOXA2 and GATA4 occupied sites were in areas of closed chromatin, devoid of histone modifications that contained varying levels of DNAme –a large proportion of which were highly methylated. We did notice that only a small proportion of FOXA2 and GATA4 sites fall in areas of H3K9me3 enrichment heterochromatin (K9-domains), though upon further investigation we find a small percentage of our endogenously compiled *cis*-regulatory regions actually reside in K9-domains indicating that heterochromatin was not the major determining factor in cell type specific pioneer factor occupancy (**Appendix S7)**. Given that the pre-existing chromatin state did not provide great discernment as to why a TF occupied certain genomic regions and not others, we next searched for motif sequences that could provide insights into alternatively expressed TFs that might modulate FOXA2s occupancy in our ectopic system (**Figure 2.10).** We found the GATA motif to be the most enriched at definitive endoderm (dEN) FOXA2 target sites that were unoccupied by FOXA2 in our ectopic system, and subsequently introduced GATA4 into the ectopic system (**Figure 2.11).** Upon co-expression of

GATA4 and FOXA2, we observed the stabilized enrichment of FOXA2 at previously only lowly enriched, sampled regions in the ectopic system indicating co-factor expression can modulate pioneer factor occupancy spectrums (**Figure 2.12).**

We continued by assessing how pioneer factor binding influences the epigenetic state with a focus on characterizing the early de-repression of silent *cis*-regulatory elements characterized by low DNA accessibility and high DNA methylated. We found that the majority of occupied regions gain some DNA accessibility, but surprisingly, only a small proportion gain significant amounts (**Figure 3.2).** The regions that become the most accessible also accumulate low levels of phased nucleosomes enriched for activating histone modifications H3K4me2 and H3K27ac, yet nevertheless we observe few transcriptional changes that we can directly associate with a FOXA2 binding site (n=82; **Appendix S9).** At FOXA2 targets with high levels of pre-existing DNAme, we observe two distinct responses to FOXA2 binding – maintenance of high methylation and dynamic loss of methylation (**Figure 3.7 and 3.8).** We find that dynamic regions have methylated CpGs closer to the enrichment summit of the FOXA2 targeted region compared to regions that remain highly methylated upon FOXA2 occupancy indicating close proximity binding of FOXA2 to methylated CpGs may induce their demethylation (**Figure 3.9).**

Loss of DNAme is thought to occur either by passive mechanisms, where CpG methylation is depleted following subsequent rounds of DNA replication or by an active mechanisms, where enzymes catalyze the removal of the methyl group through consecutive oxidation, methylated base through base excision repair (BER) or a combination of both. Thus to determine if dynamic loss of DNAme is passive, and

therefore dependent on DNA replication, we halted replication and studied the subsequent DNAme and DNA accessibility dynamics upon FOXA2 occupancy (**Figure 4.1**). In G1 arrested cells we observed similar FOXA2 occupancy and gains in DNA accessibility to non-halted cycling cells but, in contrast, we no longer observed any dynamic changes in DNA methylation suggesting that loss of DNAme is indeed dependent on DNA replication. We therefore hypothesized that FOXA2 binding in S-phase may block maintenance methylation following DNA replication which would results in the observed loss of DNAme on the newly synthesized strand. By specifically depleting FOXA2 protein levels during S-phase of the cell cycle, we observed reduced loss of DNAme at dynamic FOXA2 regions (**Figure 4.5)**. Taken together, our results provide several new molecular insights that contribute to our basic understanding of gene regulation and pave the way for a more rational use of ectopic TFs for cellular reprogramming.

5.2 Defining a pioneer factor

Originally when the pioneer factor term was devised, its definition stated that the factor was able to occupy target sites on nucleosomal DNA and subsequently remodel compacted nucleosomal arrays without the help of ATP-independent remodeling enzymes [8]. Ten years later, the same term-defining lab initially characterized the pluripotency factors, OCT4, SOX2 and KLF4, as pioneer factors for their ability to target closed chromatin DNA regions [21]. Here, they did not require the demonstration that these factors could also remodel nucleosomes upon occupancy in closed chromatin regions. Since that time, many factors have been characterized as pioneers solely for

their ability to occupy regions in closed chromatin and thus the pioneer factor definition has since morphed. It is worth reiterating that the TFs characterized as pioneer factors come from distinct TF families, have very different DNA binding domains, and likely do not function similarly. Furthermore compared to the FOXA family, there has been little investigation as to how, structurally, many of these pioneers act as such. Significant discrepancies arise even within the same class of TFs. For example, the basic Helix-Loop-Helix (bHLH) TF, ASCL1 TF has been characterized as an 'on target' pioneer factor for its ability to bind closed chromatin DNA during induced neuronal reprogramming from fibroblasts, though it cannot exert a similar function from a keratinocyte state [186]. Similarly, the bHLH TF, MYOD, has also been characterized as a pioneer factor, but can only convert specific lineages to a myotube-like phenotype and has little effect on other lineages [187]. In stark contrast to ASCL1 and MYOD, c-MYC, is a third bHLH TF not characterized as a pioneer factor and whose occupancy during iPSC reprogramming is dictated by the binding of the other reprogramming factors, OCT4, SOX2, and KLF4 [21]. Given c-MYC is a similar TF, is it then possible that c-MYC could function as a 'pioneer' in a specific situation, like ASCL1 or MYOD, and that we have just not discovered the particular scenario? These examples of bHLH TFs highlight the confusion in the pioneering definition. How do different TFs from the same family have distinct chromatin binding properties? It is possible that there is a renewed need for a more rigorous definition of a pioneer factor or a need for a more blatant hierarchical classification system of pioneer factors based on their individual properties.

5.3 OCT4 as a pioneer factor

Our findings highlight differences amongst previously described pioneer factors in their ability to ectopically access and occupy DNA target sites at high frequency. FOXA2 and GATA4 were shown to have ATP-independent remodeling capabilities *in vitro* [8] while OCT4 (along with SOX2 and KLF4 during reprogramming) have been shown to target partial motifs in nucleosomes at the initial stages of reprogramming [21,22]. We observe that FOXA2 and GATA4 demonstrate sampling at the majority of their *cis*-regulatory regions occupied in alternative cell types whereas OCT4 independently does not have this ability. In contrast, when utilizing binding data for OCT4 when it is co-expressed with the other pluripotency factors SOX2, KLF4 and cMYC [21], we do observe genomic sampling of OCT4 at human ESC target sites. Furthermore, we find the majority of ectopic FOXA2 and GATA4 binding sites to be in regions of pre-existing closed chromatin that contain varying levels of DNAme whereas ectopic OCT4 occupancy tended to occur in more pre-existing open chromatin regions. These findings were recently confirmed in mouse reprogramming where OCT4 expressed independently occupied mostly open chromatin regions whereas co-expression with the other reprogramming factors allowed occupancy of a different subset of target sites including many in closed chromatin regions [188].

These discrepancies in OCT4 independent binding compared to OCT4 binding in a reprogramming context suggest that OCT4's binding behavior is modified when it is in the presence of the other reprogramming factors. Specifically, *Oct4* and *Sox2* have a distinct and dynamic relationship that might be critical for cell specific binding patterns of OCT4. First, co-crystal structures of OCT4 and SOX2 reveal their ability to

heterodimerize at the promoter regions of pluripotency specific genes [189] and likely because of this dimerization, SOX2 and OCT4 are known to co-occupy numerous gene regulatory elements simultaneously in pluripotent cells [21,25,31]. Recent single-molecule tracking studies revealed that there may actually be a hierarchical ordering to SOX2 and OCT4 occupancy where OCT4 binding patterns are dictated by the initial occupancy of SOX2 [17]. OCT4 expression after prior SOX2 expression, demonstrated a modest increase in SOX2 DNA residence times suggesting that OCT4 may in part function by stabilizing SOX2 occupancy. In contrast, SOX2 expression after previous OCT4 expression, resulted in significantly decreased OCT4 target search times suggesting that SOX2 might be needed for OCT4 to ultimately find and occupy its target sites [17]. Taken together, this suggests that OCT4 binding or its pioneering activity may have some dependency on SOX2.

The presence of the other reprogramming factors in conjunction with OCT4 might result in distinct post-translational modifications (PTM) to the OCT4 protein, which do not occur when it is expressed independently and could explain the distinct occupancy patterns we observe in our system. Recent studies have demonstrated that OCT4 PTMs can vary in a cell type specific context. Purified, recombinant OCT4 protein was subsequently incubated with a variety of distinct cellular extracts. The resulting PTMs were profiled by Mass Spectrometry and distinct phosphorylation patterns in the POU, DNA binding domain that are predicted to impact OCT4 protein and DNA interactions were observed based on cellular context [190]. Furthermore, removal of another PTM, O-GlcNAcylation (O-GlcNAc), present on both OCT4 and SOX2 proteins in pluripotent conditions and during reprogramming in MEFs, results in inefficient iPSC

reprogramming as well as loss of pluripotency and self renewal in ESCs [191,192]. O-GlcNAc modified SOX2 is critical for its protein-protein interactions [193] as SOX2 O-GlcNAc deficient protein occupies unique genomic regions that do not contain OCT4 motifs [193]. Overall, it is clear that PTM modification can alter TF binding spectrums and that distinct cellular context can result in different modifications. It should be interesting to examine the PTM spectrum that OCT4 acquires when it is expressed by itself or with the reprogramming factors to determine if these distinctions can explain the differential occupancy patterns demonstrated for OCT4 in our study.

5.4 Cell type specific occupancy spectrum even among of pioneer factors

A study examining the occupancy patterns of FOXA1 across three different breast cancer cell lines revealed that despite the pioneer properties of FOXA1, it still displayed some cell type specific binding patterns [16]. These initial findings were accomplished using ChIP followed by microarray profiling as opposed to high throughput sequencing and thought to be of low resolution. Yet our study confirmed the speculation that pioneer factors display distinct and specific binding patterns across various cell types, even under super-physiological expression conditions (**Figure 2.1 and 2.3**). Initially, this result was surprising given the unique chromatin binding capabilities of FOXA proteins, however studies of ectopic pioneer factor expression in subsequent years have revealed H3K9me3 heterochromatin restrictions to ectopic pioneer factor occupancy [21]. Nevertheless, our method of using an ectopic expression system to study pioneer factor occupancy at endogenously occupied target sites in alternative cell types revealed a potentially unique quality that pioneer factors possess –

the ability to sample most of their motif containing *cis*-regulatory targets (**Figure 2.4 and Figure 2.5).** We subsequently demonstrated that a subset of sampled target sites can be converted to high frequency occupied regions by the expression of additional TFs whose co-occurring motifs are present at these regions (**Figure 2.12).**

However, we are still only able to convert a subset of the predicted target sites. Further motif analysis revealed the possibility that multiple TFs may play a role in stabilizing FOXA2 enrichment at these target sites. We find that sites that become occupied by FOXA2 in the presence of GATA4 do not gain DNA accessibility. This experiment indicates that cooperativity of these two factors is not predicated on the recruitment of chromatin remodeling machinery to alter accessibility and allow for target site access. Instead, it appears that both factors are enriched at similar genomic regions in closed chromatin. GATA4 is also sometimes considered a pioneer factor and is able to occupy this subset of targeted regions when it is expressed independently and its presence may cause the stable accumulation of FOXA2 enrichment at these locations. It is also possible that these FOXA2 sites are actually indirect target sites that occur through protein-protein interaction with GATA4 directly contacting the DNA. ChIP utilizes formaldehyde crosslinking, which links both protein-DNA interactions as well as protein-protein interactions. It is therefore challenging to discern if occupied regions, as assessed by ChIP-seq, are the result of direct protein-DNA interaction or through indirect, protein-protein interactions. However, most of these target sites contain motifs for both FOXA and GATA indicating that both factors may directly bind DNA. Furthermore, when browsing these regions using the Integrated Genome Viewer (IGV), we often observe that accumulation of GATA4 and FOXA2 are not precisely

overlapping. Instead, it appears that summits of the factors are slightly shifted away from the other (GATA4 left shifted, FOXA2 right shifted **Figure 2.12D)**. However, it is worth noting that we used DNA sonication based fragmentation of the DNA prior to immunoprecipitation, which results in less precise genome-wide mapping of TFs and histone modifications. Instead, ChIP-exonuclease experiments that map more precise, minimal binding locations for these factors under these conditions might provide better insights.

The idea of redistribution of pioneer factor occupancy by distinct cellular influences has recently been discussed in the literature but with differing views on how it is accomplished. First, a subset of SOX2 binding sites that have suboptimal chromatin state and motif positioning in mouse ESCs have recently been shown to be dependent on the concurrent binding of PARP1 and are subsequently lost when the PARP1 gene is knocked out [19]. Through *in vitro* nucleosome binding assays, DNase I footprinting and biochemical studies, the authors speculate that PARP1 binding at these regions may reduce the minor groove width which allows for SOX to better recognize its suboptimal motif [19]. A second study demonstrated that TNF-alpha treatment of breast cancer cells results in a new subset of FOXA1 target sites. The binding and pioneering ability of FOXA1 is still required at these targets to allow subsequent access of the estrogen receptor [20]. While TNF-alpha stimulation also activates NF-kappaB genomic binding, the authors do not observe complete overlap in FOXA1 and NF-kappaB occupancy at the new subset of FOXA1 targets indicating that NF-kappaB alone is not responsible for the redistribution of FOXA1 targets. Instead, the authors speculate that TNF-alpha signaling may cause a post-translational modification to FOXA1 that alters

its binding capabilities, though no data are shown to validate this claim. Lastly, a third study demonstrated that estrogen receptor (ER) activation in breast cancer cells redirects FOXA1 occupancy to a new and very small subset of target sites further adding to the complexity. Here, the authors speculate that at these targets, ER attains pioneering activity via its recruitment of chromatin remodeling machinery to these loci though they do not demonstrate that fact (see Chapter 1 for further description) [18]. Regardless of the mechanism, it is clear that cell state can influence even pioneer factor occupancy. Though it is worth mentioning that in all these studies, as well as in our own study, pioneer factor occupancy is only minimally redistributed and that the majority or pioneer factor target sites remain stable.

There are certainly other factors we have not considered in our study of cell-type specific pioneer factor occupancy. As stated above for OCT4, FOXA2 may have distinct PTMs that alter its interaction spectrum and ultimate occupancy profile across cell types. Though few PTMs for FOXA2 have thus far been identified and these have mostly been implicated in protein stability, not DNA binding [194,195]. While we have considered the pre-existing chromatin state of the cell, we have only looked at a subset of known histone modifications. The marks that we considered are thought to be the most consequential based on Roadmap Encode guidelines [196] and our own experience, but this does not exclude the fact that a less characterized modification may also influence TF occupancy. Finally, we have not considered the three-dimensional architecture of the genome and its influence on TF binding.

5.5 Limited influence of pioneer factors to significantly remodel chromatin

Our data demonstrated that FOXA2 occupancy at pre-existing closed chromatin regions results in overall minimal gains in ATAC-seq representative DNA accessibility with only a small number of regions gaining significant amounts of accessibility (**Figure 3.2)**. The overall minimal gain in ATAC-seq signal observed at most closed chromatin regions might result from an intrinsic difference between the association of DNA and a nucleosome once it is occupied by a TF compared to its non-occupied state (**Figure 3.3).** We initially considered regions that gained significant ATAC-seq signal, nucleosome depleted, though recent studies suggest that DNA accessibility can result without great changes in nucleosome occupancy (see Chapter 1) and our current data cannot distinguish between these two possibilities [3,20,90,91]. It is possible that the low MNase digestion conditions used in these studies capture low frequency nucleosomes competing for the same DNA binding location with a TF as the process is likely quite dynamic given quick DNA residence times for TFs [5,17,197]. Indeed, we do observe the accumulation of phased, modified nucleosomes at sites that gain significant DNA accessibility indicating potential nucleosome eviction in at least some proportion of cells.

The surprising result that not all FOXA2 occupied regions gain significant DNA accessibility prompted us to examine differences between the regions that do gain and those that do not gain significant DNA accessibility. First, recent *in vivo* data confirmed previous *in vitro* binding observations, that FOXA occupancy only results in remodeling of nucleosomal arrays when they are compacted with histone protein H1 and that there is little effect on non-compacted arrays [3,8]. When we examine FOXA2 occupied regions in pre-existing closed chromatin regions, we define closed chromatin as any regions

with ATAC-seq of RPKM < 1. Whether these regions are compacted with H1 protein or not, we cannot determine from ATAC-seq alone. A better understanding of how ATAC-seq and MNase data correspond would be helpful in assessing these differences. Furthermore, mapping of histone protein H1 would give a better indication of how much of this 'closed chromatin' is actually compacted by H1. We do observe the enrichment of FOXA motifs at regions that gain significant accessibility compared to regions that do not and observe a wider distribution of the motif surrounding the peak summits (**Figure 3.4**). We speculate that motif positioning near the nucleosome dyad may be critical for chromatin remodeling as the motif position has been shown to be critical for FOXA1 remodeling in vitro [160]. Furthermore, our speculations mainly focus on the ATP-independent chromatin remodeling imposed by FOXA2, but it is possible that there is some recruitment of chromatin remodeling machinery to these loci allowing for significant increases in ATAC-seq signal accumulation.

5.6 Loss of DNAme at a subset of targeted regions

We observe two distinct outcomes after FOXA2 binding at highly methylated regions – regions that lose most DNAme and regions that retain high levels (**Figure 3.6**). The core FOX motif is a six base pair sequence (TAAA(T/C)A) that itself does not contain a CpG, but its immediate flanking sequence may in fact contain a CpG. Co-crystal structures of DNA binding domain of FOXA with DNA though have demonstrated that FOXA actually occupies around 14 base pairs of DNA sequence which would allow for FOXA2 to physically interact with CpG methylation [137]. However, CpG methylation even outside of the core and flanking motif sequence has also been shown to impact TF

binding for other pioneer factors. For instance, as DNAme increases nucleosome stability [198], its presence within 100 base pairs of an OCT4 binding site was shown to restrict OCT4 occupancy and subsequent nucleosome remodeling [199]. In contrast to the influence DNAme has on OCT4 occupancy, our results indicate that while FOXA2 can occupy highly methylated regions regardless of CpG location, its binding can actually influence subsequent DNAme levels when the CpG is closest to the motif sequence (**Figure 3.8**). .

It was surprising that we observed regions that remain highly methylated after FOXA2 occupancy as a general correlation with FOXA occupancy and low DNAme has been previously described [24]. It is possible that more precise refinement of the FOXA2 occupied regions via assays such as ChIP exonuclease would determine if FOXA2 itself physically interacts with DNAme or if the DNAme is just within the surrounding region. Though when we refine our analysis and only look at 20 base pairs surrounding our peak summit, we still observe ~15% of regions contain highly methylated CpGs. In addition, *in vitro* DNA binding studies demonstrate that Forkhead factors can associate with methylated DNA fragments [200]. We initially observed a broad loss of DNAme at the majority of FOXA2 target sites during the differentiation from human ESCs to dEN [25]. Upon closer inspection of this data we actually find n= 217 regions that retain methylation within 50 base pairs of the FOXA2 occupied region in dEN indicating that even under physiological conditions, FOXA occupies regions with proximal methylation on the DNA (**Figure 5.1).** Though the majority of the highly methylated regions targeted by FOXA2 in dEN lose DNAme (n = 2550).

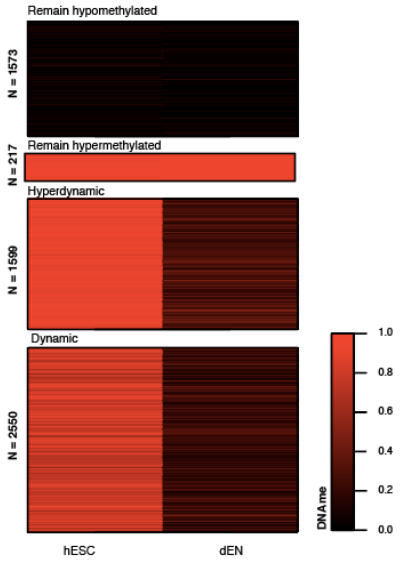


Figure 5.1:
DNAme levels in human ESCs and dEN of regions bound by FOXA2 in dEN. DNAme in the human ESC condition is measured by WGBS of matching CpGs in the dEN FOXA2 ChIP-BS-seq experiment. Regions remain unmethylated after binding, remain methylated after binding or dynamically lose DNAme. Hyperdynamic regions are a subset of the Dynamic regions with initially high DNAme levels.

We know that the majority of regions bound by FOXA2 in dEN, do become functional *cis*-regulatory elements later in development [25]. For example, we observe FOXA2 occupancy and loss of DNAme at *cis*-regulatory elements in the Albumin enhancer locus, though these genes are only activated later in hepatocyte development. Looking back at DNAme dynamics in our ectopic system, it remains to be seen whether the *cis*-regulatory elements that lose DNAme are in fact functional regulatory elements in development. To get at this question, we overlapped the dynamic set (class 3) with a compiled list of functional *cis*-regulatory elements across all Encode cell types and find that actually n= 3280/9111 regions that display dynamic loss of DNAme in our ectopic system are putative functional and active *cis*-regulatory elements in at least one other cell type. Surprisingly though, a similar analysis finds that 5740/8792 regions that

remain methylated (class 2) also appear to overlap with putative active regulatory elements in at least one other cell types. This indicates that in general ectopic FOXA2 targeting, even within regions that remain methylated, is mainly occurring at putative *cis*-regulatory elements. The increased distance from the FOXA2 peak summit to the nearest CpG at class 2 target sites may indicate that either other DNA binding factors are needed at these *cis*-regulatory elements to induce the loss of DNAme or that the methylation changes are not needed or relevant at that particular stage. Overall while few dynamic elements are associated with a transcriptional change, the loss of DNAme at these regions represents the initial step toward the activation of these loci and thus is critical for enhancer competence and proper development [2].

5.7 Replication dependence on DNA demethylation

To assess if the dynamic loss of DNAme observed at a subset of FOXA2 binding sites is dependent on DNA replication, we generated a system in which we were able to halt the cell cycle, subsequently induce FOXA2 and perform ChIP for the factor in the G1 arrested condition (**Figure 4.1**). Indeed, we observed that the loss of DNAme in dynamic FOXA2 target sites is dependent on DNA replication (**Figure 4.3**), suggesting a passive contribution rather than a completely active demethylation. Thus we speculated that there are two possible explanations for the observed dependence on DNA replication. First, we thought that FOXA2 binding during G1 of the cell cycle might mark or alter the methylated cytosine bases (i.e. with 5hmC) making them less recognizable to the maintenance methylation machinery and subsequently passed over by DNMT1 [126,127]. 5hmC is a by-product of an oxidation reaction catalyzed by TET enzymes on

5mC and is either subsequently not maintained by DNMT1 due to its lack of affinity for 5hmC or removed from the DNA through BER (see **Chapter 1**). Importantly, during bisulfite conversion experiments, 5hmC and 5mC react the same way and both appear as 'methylated'. Thus the possibility remains that we characterized the high levels of DNAme at FOXA2 target sites we observed when the cell cycle is halted in G1 as 5mC when it is actually 5hmC. However we had good reason not pursue this line of thought more thoroughly. First, the BJ fibroblasts in which we used for our experiments have very little expression of any of the TET enzymes, and fibroblasts have been used as a negative control in 5hmC experiments for their lack of 5hmC [201]. FOXA2 induction does cause the increase of TET2 from FPKM of 2.5 to FPKM of 6, which might be sufficient for low levels of 5hmC accumulations. However, TET2 recognizes unmethylated cytosine bases via its interaction with a CXXC domain-containing co-factor IDEX [202], and we do not detect any expression of the IDEX gene following FOXA2 induction. While this information provides evidence that enzymatic reactions caused by the TET enzymes are unlikely to induce 5hmC in our system, it does not exclude the fact the modification could be catalyzed at these regions by other means. Furthermore, a recent publication demonstrated an interaction between FOXA1 and DNA repair enzymes quite convincingly [185], though the authors do not speculate as to the potential demethylating enzyme, examine subsequent accumulations of 5hmC or test dependence of loss of DNAme on replication.

Secondly, we suspected that S-phase binding of FOXA2 following DNA replication might physically block the maintenance methyltransferase activity from re-methylating nascent CpGs and established a system, which depleted FOXA2 protein

specifically during S-phase of the cell cycle (**Figure 4.4**). We hypothesized that this depletion would limit the dynamic change in DNAme observed at FOXA2 occupied regions. Indeed, even with low levels of FOXA2 protein remaining during S-phase of the cell cycle, we surprisingly observed less dynamic loss of DNAme at FOXA2 target sites (**Figure 4.5**). With this data, we suspect that S-phase binding of FOXA2 at hemi-methylated DNA target sites, blocks maintenance methyltransferase activity. An S-phase FOXA2 ChIP would be insufficient to demonstrate this point unquestionably as there are multiple origins of replication firing at once and FOXA2 will likely remain bound at some regions that have yet to undergo replication further complicating the results. To demonstrate this definitively, we need to perform FOXA2 ChIP-BS ensuring just one round of DNA replication has taken place, while subsequently using a hairpin adapter to capture equivalent information from both DNA strands while making the next generation sequencing libraries.

DNMT1 is the methyltransferase enzyme that recognizes hemimethylated CpG dinucleotides following replication and catalyzes DNAme on nascent strands. Re-methylation is thought to occur rapidly after DNA replication because of DNMT1's co-localization with the replication fork through its interactions with PCNA and UHRF1 which position DNMT1 at hemi-methylated DNA [92]. However, recent evidence from our lab (that is currently under review for publication) indicates there is actually a lag in DNAme following DNA replication that is gradually consolidated over time (up to 4 hours following replication; **Figure 5.2***) [203]. As FOXA2 must be released from the DNA to allow the replication fork to pass, the subsequent binding of FOXA2 might actually be able to occur prior to nascent strand methylation. This mechanism would result in

passive loss of methylation that does not require TET activity and would support our current findings.

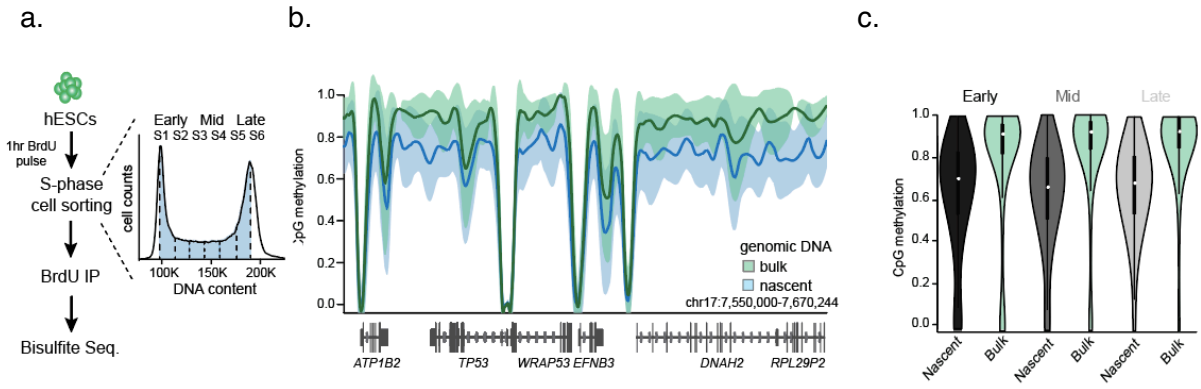


Figure 5.2:* (figure submitted for publication elsewhere [204])

a. Schematic representation of BrdU-IP-BS-seq experiment to capture nascent DNA
b. IGV browser view of methylation level in nascent DNA captured in S-phase compared to bulk cell cycle genomic DNA
c. Mean methylation values of nascent DNA at Early, Mid and Late points during S-phase of cell compared to bulk genomic DNA capture for equivalent stages

There is little data examining transcription factor recruitment to DNA during S-phase, post-replication. One paper demonstrated through early S-phase, late S-phase and early M-phase ChIP that Cohesin remains bound to similar locations, and this observation was subsequently verified by nascent chromatin capture and protein purification studies [30,205]. Likewise, CTCF has been shown to associate with nascent DNA molecules [16]. A recent paper uses a single cell imaging approach that utilizes EdU (a thymidine analog) incorporation into nascent DNA strands coupled with a proximity ligation assay reaction that only occur if two antibodies (one against EdU and one against another factor) are in close vicinity [206]. By this method, the authors demonstrate that replication may be needed to recruit certain TFs (like FOXA2) to DNA during differentiation as blocking synthesis results in minimal TF recruitment to DNA [206]. Furthermore, they demonstrate a lag in subsequent enrichment of H3K27me3 occurs

post- replication and suggest that this lag may allow for a 'window of opportunity' for TFs to access target sites before the modification of newly synthesized histones leads to chromatin restriction [206]. Though the lag in a repressive epigenetic modification is similar to the lag in DNAme that our lab observes (see **Figure 5.1**), the 'window of opportunity' mechanism proposed for TF occupancy as a result of this lag is in distinct contrast to what we propose. We postulate that S-phase binding of FOXA2 actually causes the reprogramming of DNAme patterns as we demonstrate that FOXA2 can bind to both methylated and unmethylated DNA substrates without preference (**Figure 4.2**). S-phase binding of FOXA2 during the lag in DNAme following replication would only provide a mechanism of blocking DNMT1 maintenance methylation. A pulse of EdU to mark nascent DNA strands prior to FOXA2/EdU sequential ChIP-BS-seq would be an ideal experiment to demonstrate this hypothesis, yet the limiting DNA amounts acquired from the FOXA2 ChIP make success unlikely. Instead, we are working on a time course experiments where we will mark nascent DNA strands with varying pulses of EdU, immunoprecipitate FOXA2 and run a dot blot for captured EdU to demonstrate FOXA2 occupancy on nascent DNA. While these experiments will not be able to assess the lag in DNAme, they will demonstrate, in a global way, how quickly FOXA2 can associate with nascent DNA.

Because of the similarity in DNA binding domain structure of FOXA2 to the linker histone H1 [137], it was interesting to understand how linker histones are inherited post-replication. Following DNA replication, parental histones proteins along with newly synthesized histone proteins are quickly reassembled on both DNA strands via replication-dependent histone chaperone proteins [207]. Much of the literature focuses

exclusively on inheritance of core histone protein H2A/B, and H3/H4, though it appears that prior to replication there is a large synthesis of non-polyadenylated core and linker histone transcripts indicating an equivalent need for all histones during S-phase [207,208]. Furthermore, studies of histone incorporation post- replication that observed linker histones dynamics, demonstrate that H3/H4 proteins are first incorporated, followed by H2A/H2B with histone H1 the last to be assembled once the nucleosome is intact - though this all occurs within a short time (minutes to hours depending on the study) following the passage of the replication fork [209-211]. Nascent Chromatin Capture followed by protein purification though indicates that linker histone H1s are mostly associated with a more mature chromatin species (2 hours post-replication) compared with immediate nascent chromatin and this may be due to the acetylation of newly synthesized H4 histones [205]. Taken together, these results indicate that histone H1 can potentially be incorporated to new DNA shortly (minutes to hours) after passing of the replication fork, after the assembly of the core histone particles. Given the structural similarities between FOXA and linker histones, it is possibly that FOXA follows a similar mechanism and can incorporate into nascent DNA following the assembly of nucleosomes. A time course following EDU incorporation and nascent chromatin capture to examine FOXA2 association with nascent chromatin might elucidate the precise timing of FOXA2 recruitment to nascent DNA.

5.8 Future directions

This work provides various intriguing paths to further explore pioneer factor limitations, capabilities and mechanisms. Further work could be dedicated to

investigating the 'sampling' phenotype that we describe for FOXA2 and GATA4, across various other TFs to determine the uniqueness of this phenotype and to identify potential new factors that may have undiscovered pioneer capabilities. Because we observe that sampled sites are more frequently stabilized by co-factor expression, one could investigate these genomic features more extensively utilizing a combination of genome-editing techniques [212] and parallel reporter/binding assays [213]. First, to establish that the observed low-enrichment at sampled target sites is specific and a result of motif sequence, FOXA2 motifs could be deleted from a chosen set of sampled regions and subsequent local ChIP experiments could determine if sampling is lost at these regions as a result. Likewise, motifs of co-factors could be deleted at similar subsets of sampled regions to demonstrate the stabilization of FOXA2 targets by co-factors is dictated by motif occurrence of co-factors such as GATA4. These experiments would also provide insights into direct and indirect TF binding. For instance, upon removing the FOXA2 motif from a particular locus, if after co-expression with GATA4 we still observe the stabilized enrichment of FOXA2 binding, this would indicate FOXA2 enrichment is likely due to the indirect tethering of FOXA2 to the GATA4 protein, and not the co-localized binding of both factors to DNA.

In parallel, to examine a subset of sampled target sites in a more high-throughput way, one could utilize current massive parallel reporter assay (MPRA) strategies [214] for TF occupancy instead of luciferase expression. MPRA oligomers could be designed, utilizing a mutagenesis strategy at *in vivo* FOXA and GATA4 motif sites for a subset of regions, and stably integrated into the BJ fibroblast genome (along with other cell types) using lenti-viral MPRA constructs. Subsequent, expression and/or co-expression of

both TFs would be performed followed by ChIP and amplification of just the MPRA fragments to determine how mutagenesis of FOXA and GATA motifs affected sampling and high enrichment binding at these regions. Here, utilizing an integrated MPRA system would be key as transient transfection of MPRA oligomers would not induce chromatinization of the fragments [213]. Even though we hypothesize that chromatin structure has less influence on pioneer factor binding, these strategies would allow us to assess specific manipulations of co-factors and sequence features influence on pioneer factor binding under more physiological conditions.

It would be interesting to explore some of our findings about pioneer factor occupancy and cooperativity in physiological relevant systems to determine their importance. For instance, our lab previously identified an interesting *cis*-regulatory element that appears to be uniquely regulated in the definitive endoderm (dEN) germ layer and is co-bound by FOXA2 and GATA4 among other factors in dEN (**Figure 5.3**) [25,179]. This region is highly methylated in human ESCs, definitive ectoderm (dEC) and definitive mesoderm (dME), and subsequently only loses DNAme upon dEN specification [25]. In addition to FOXA2 and GATA4 occupancy at this region in dEN, we also observed FOXA1, OTX2, SOX17, EOMES, and GATA6 occupancy indicating that this is may be a critical regulatory element in dEN [179]. When we examine this region in our ectopic system, we find that the region has low levels of DNAme in BJ fibroblasts, yet is still not occupied by FOXA2 (**Figure 5.3**).

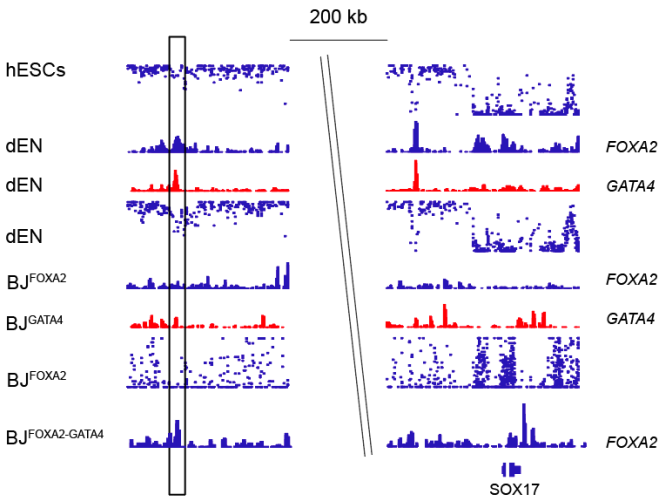


Figure 5.3:
IGV browser shot of the SOX17 locus and potential *cis*-regulatory region 200kB upstream (boxed region). Top track displays DNAme in human ESCs from 0 -100%. Next two tracks display FOXA2 and GATA4 ChIP in dEN. Forth track shows DNAme in dEN from 0 -100%. Next two tracks display FOXA2 and GATA4 ChIP in BJ fibroblasts. Final track shows FOXA2 ChIP in BJ fibroblasts that co-express both FOXA2 and GATA4. All ChIP tracks are .tdf files normalized in IGV for coverage and displayed from 0-2.

Upon co-expression of GATA4 with FOXA2 in BJ fibroblasts, we begin to see low-enrichment accumulations of both FOXA2 and GATA4 at this region indicating that cooperative binding is associated with TF occupancy at this locus. To determine how important (if at all) the cooperativity of FOXA2 and GATA4 is at this locus, it would be interesting to individually knock out the FOXA and GATA motifs at this region in human ESCs using CRISPR. Subsequent differentiation experiments to dEN would allow us to assess if either FOXA or GATA were critical for the activation of this *cis*-regulatory element. Subsequent ChIP for this region would determine if any of the other endodermal TFs were able to occupy this region in the absence of either the GATA or FOXA motif. In line with this, methylation profiling would determine if these factors were sufficient for the loss of DNAme observed at this region and if this *cis*-regulatory

element is critical for dEN state. Utilizing genome-editing strategies in human ESCs would allow us to assess the relevance of our ectopic system findings in a developmental model system.

Further outside the scope of direct follow up experiments related to my thesis work, it would also be interesting to now utilize some of the acquired techniques and our specific findings to assess FOXA's role in development and in disease. First, understanding more about the pioneering role of FOXA in establishing critical gene regulatory networks during the early stages of development would be an interesting pursuit given the advent of single cell sequencing technologies. We know that FOXA2 is first expressed in the epiblast and is critical for early specification as knock out mouse models are embryonic lethal due to axis abnormalities, and absence of node and notochord structures [10,11]. Other than gross anatomical characterizations of knockout and conditional mouse models, there is little understanding of the molecular regulation imposed by FOXA2 on critical gene regulatory networks in early development. Furthermore, FOXA1 and FOXA2 are generally thought to have quite similar molecular features yet neither factor can fully compensate for the other in knockout animals indicating each has specific developmental roles. Our lab currently has the ability to rapidly and efficiently generate gene knockouts by injecting Cas9 and guide RNAs into fertilized embryos and following their subsequent early differentiation, blastocysts are embedded into pseudo-pregnant females that can later be isolated for characterization of particular developmental stages. Utilizing this kind of knock out strategy would allow us to isolate FOXA1 or 2 knockout embryos at early expression (E6.5) and later developmental time points (E9.5), and perform single cell RNA sequencing assays.

With this, we would be able to classify the distinct cell populations and proportions that are able to form in comparison to wild type conditions. Furthermore, bulk and single-cell ATAC-seq and methlyome analysis could be performed in parallel to map the critical regulatory elements that fail to become properly remodeled in the absence of FOXA proteins. This would provide a molecular view of the gene regulatory pathways modulated by FOXA2 during *in vivo* early development.

Next, focusing more on the role of FOXA2 and disease would be interesting to pursue. As FOXA is expressed in most adult endodermal tissues, much of the focus in FOXA disease research has been investigating its pioneering role in defining steroid receptor binding patterns across multiple cancers such as breast, prostate, and thyroid [215]. FOXA though, is also expressed in midbrain, dopaminergic neurons (mDA) and its down-regulation may play a role in the development of Parkinson's Disease (PD) [216,217]. Indeed, conditional knockout mouse models for FOXA demonstrate that mDA lose dopaminergic properties [217]. Furthermore, PD mDA neurons contain aberrant methylation patterns compared to control mDA neurons where hypermethylated DMRs were localized mainly to low CpG density enhancer regions that are often occupied by TFs such as FOXA in other cell types [218]. This data suggests that FOXA occupancy in mDA neurons may in fact, be critical for the loss of DNAme at *cis*-regulatory regions that function in essential regulatory networks dictating dopaminergic cell state. Due to the limited cell number of *in vivo* mDA neurons in mice, little work has been done in characterizing the TF networks controlling these cells. Recently, differentiation protocols from human ESCs have become more robust [219,220] allowing for the development of a human model system in which to study PD better. Utilizing a human

ESC differentiation system would allow us to specifically assess the role of FOXA proteins and loss of DNAme during mDA commitment as well as in the establish cell type.

As described above, the definition of a pioneer factor has morphed over the years, making a complete mechanistic understanding of pioneer factors as an entire class of proteins unlikely due to their distinct domain structures. Mechanistic studies into individual pioneer factors, like this thesis, will allow for a greater understanding of the distinct occupancy patterns and chromatin remodeling capabilities through which particular pioneer factors initiate responses that ultimately alter the cellular state. Establishing the underlying features that contribute to pioneer TF recruitment and enhancer assembly will allow for the engineering of novel transcriptional changes that utilize pioneer TFs to override the preexisting chromatin landscape within an ectopic cellular environment and may ultimately lead to more effective and efficient reprogramming and differentiation.

Chapter 6.
Materials and Methods

Parts of this chapter are submitted for publication elsewhere[1]

Cell culture

Clonal FOXA2 doxycycline inducible cells lines were derived from an immortalized BJ foreskin fibroblast cell line from ATCC (BJ-5ta; CRL-4001). Cells were cultured in MEM-Alpha (Life technologies: 32561-037) with 10% FBS, 1% pen-strep, .01mg/mL hygromycin B and 5ng/mL bFGF. Derived BJ^{FOXA2} lines were grown in the same conditions plus 0.5ug/mL Puromycin.

BJ cell line generation

Cells were infected with pTRIPZ-FOXA2, pTRIPZ-RFP, pTRIPZ-POU5F1 at an MOI ~1. Following infection, cells were selected with Puromycin (1ug/mL) and replated at a high dilution to ensure separation for clonal expansion and isolation. After two weeks of growth, individual clones were picked, expanded and screened. Criteria for inclusion in the current study included uniform expression of FOXA2 and minimum basal FOXA2 expression. Clones were maintained in .5ug/mL puromycin containing media following expansion. To induce FOXA2, doxycycline was added at .5ug/mL.

Cloning and Constructs

To generate pTRIPZ-FOXA2, RFP, FOXA2-CDT1, pTRIPZ-POU5F1, pTRIPZ inducible lentiviral vector (Thermo Scientific) and full-length *FOXA2* were assembled using Gibson Assembly® Master Mix (NEB). pTRIPZ empty vector was digested with XhoI and MluI to remove shRNAmir regulatory sequences, and digested ends were blunted. The linearized pTRIPZ backbone was digested with BsiWI to generate two fragments,

each with one sticky end. The fragments were gel extracted, purified, and ligated using the Quick LigationTM Kit (NEB). Primer sequences are listed in Table 1.
To generate HaloTagged-FOXA2 construct, full-length FOXA2 was ligated to pFN21A (Promega)

GATA4-V5 and POU5F1-V5 constructs were obtained from the Broad Institutes Genomics Perturbations platform and are available to purchase through Thermo Fischer.

Protein purification

293Ts were transfected with pFN21A-FOXA2. Purification was completed following Promegas's Halotag Protein Purification System. Briefly, 48 hours following transfection, cells were harvested, lysed and gently sonicated four times on Branson Sonifier at 10% amplitude for 15 seconds. Sample was diluted 1:3 with protein purification buffer (1X PBS, 1mM DTT, .0005% NP-40) and centrifuged to remove debris. Halo-Resin was washed in purification buffer, added to lysate and incubated at 4C overnight. After incubation resin was washed and FOXA2 protein was cleaved via the addition of TEV-protease during an over night incubation at 4C. Purified protein was assessed via commassie blue gel and western blot.

EMSA

EMSA was performed using LightShift Chemiluminescent EMSA kit (Pierce). Purified halotagged-FOXA2 protein (3-6ug) was mixed with duplexed, biotinlyated probes

(20fmol/ul) without competitor DNA. Unlabelled probes (non biotinlyated) were added at 10X-100X concentration of biotinlyated probes. Binding reactions were incubated for 20 minutes at room temperature before loading onto a 6% DNA retention gel (Invitrogen). Complexes were transferred to nylon membrane (Invitrogen) and crosslinked via UV radiation in Statalinker. Biotinlyated DNA was detected by chemiluminescence.

Chomatin Immuoprecipitation

Cells were crosslinked with 1% formaldehyde for 10 minutes at room temperature, and quenched with 125mM glycine at room temperature. Chips were performed as previously described [25] by isolating nuclei and shearing DNA to 200-600 basepair fragments using Branson sonicator. Antibody incubation with chromatin was performed overnight. ~10 million cells were used per FOXA2 ChIP with 1ug of antibody/ million cells. ~1 million cells were used for each histone ChIP. Following an overnight incubation, antibody-protein complexes were isolated using Protein G/A beads (Life Technologies) and sequencing libraries were generated. Libraries were generated as previously described [25,221] and Libraries were sequenced on a HiSeq 2500 at 11pmol.

Chomatin Immuoprecipitation Bisulfite Sequencing

To generate bisulfite converted DNA libraries following ChIP, we used Nugen Ovation UltraLow Methyl-seq Kit (0335-0336). Bisulfite conversion was performed with EpiTect Bisulfite kit (Qiagen) with carrier DNA. Libraries were sequenced on a HiSeq 2500 8pmol with 35% PhiX spike in.

Antibodies

Chips were performed using: FOXA2 (R&D: AF2400), H3K4me2 (ActivMotif: 39141), H3k27Ac (ActivMotif: 39133), H3K27me3 (ActivMotif: 39155), V5 (MBL: M167-3), OCT4 (Active Motif: 39811)

Immunostaining performed with the following antibodies: FOXA2 (R&D: AF2400), V5 (MBL: M167-3), OCT4 (Cell Signaling: 2750)

Westerns: FOXA2 (R&D: AF2400), V5 (MBL: M167-3), H3 (Abcam: ab1791), OCT4 (Cell Signaling: 2750)

IGV Browser shots

All browser shots were created in Illustrator by exporting .svg files from Integrated Genome Viewer (IGV). Data were imported into IGV as normalized TDF files and scaled to the same values (2). Genomic location is listed in figure legend.

Whole genome bisulfite sequencing

WGBS was performed as previously described using Swift Acell-NGS Methyl-Seq DNA kit.

ATAC sequencing

Tagmentation was performed on whole nuclei at 37C for 45 minutes as previously described in [176]. DNA was isolated on PCR min-elute columns (Qiagen) and a small amount of the DNA was amplified for 9, 12 and 15 cycles to determine optimal cycling conditions. The rest of the DNA was then amplified using the chosen cycle number and

PCR libraries were purified using double sided Ampure clean up to remove high molecular weight fragments. .55x Ampure volume was added to the PCR, mixed and incubated. Supernatant was removed following magnet separation and cleaned-up with a 1X Ampure volume. Libraries were sequenced on a HiSeq 2500 at 8pmol.

RNA sequencing

RNA was isolated with RNeasy columns (Qiagen) and non-stranded libraries were performed using Illumina's standard Tru-Seq kit. Libraries were sequenced on a HiSeq 2500 at 11pmol.

RTq-PCR

cDNA synthesis was performed with 600-2000 ng of RNA using the RevertAid TM First Strand cDNA Synthesis Kit (Thermo Scientific) with oligo(dT)18 primer. Quantitative PCR (qPCR) primers were designed with Primer-BLAST (NCBI). Primers were designed to span an exon-exon junction, amplify 70-200 bp of cDNA, and amplify all isoforms of a transcript. qPCR was performed with 3-4 technical replicates using a 1:100 or 1:1000 dilution of cDNA, Power SYBR Green Master Mix (Applied Biosystems) and 500 nM of forward and reverse primers on the ViiATM 7 Real-Time PCR System (Applied Biosystems). *ACTB* and *HPRT1* were used as endogenous controls. Relative gene expression was calculated with the comparative CT (ΔΔCT) method using ExpressionSuite Software v1.0.3 (Biosystems).

Western

Nuclear proteins were extracted in standard RIPA buffer supplemented with protease inhibitors (Roche). Equal amounts of extracts were mixed with LDS (Life Technologies) and BME and boiled at 95C for 5 minutes. Samples were loaded onto an NuPage Novex 4-12% Bis-Tris gel (Life technologies) and electrophoresed for 1 hour at 200 volts in 1X MES buffer (Life Technologies). Proteins were transferred to PVDF membrane via iBlot transfer system (Life technologies). Membranes were blocked in 5% Milk/TBST for 1 hour at room temperature and membranes were incubated with primary antibodies in 5% Milk/TBST over night at 4C. FOXA2 primary antibody was diluted at 1:3000 and H3 primary antibody was diluted at 1:10,000. Membranes were washed and incubated in secondary antibodies in TBST at 1:10,000 dilution. Detection was performed with SuperSignalTM West Dura Chemiluminescent Substrate (Thermo Scientific).

Immunostaining

Cells were fixed in 4% formaldehyde for 15 minutes at room temperature. After washing, permeabilization and blocking was performed with 4% FBS/0.4% Triton in PBS for 1 hour at room temperature. Primary antibody staining was performed with 2% FBS/0.2% Triton in PBS overnight at 4°C. Secondary staining was performed with a fluorophore-conjugated antibodies in PBS for 1 hour at room temperature.

Cell cycle arrest and FACs analysis

Cells were halted in G1 by the addition of 500mM Mimosine (Sigma) treatment over night.

Cell proliferation was determined using the Click-iT® Plus EdU Flow Cytometry Assay Kit (Life Technologies). 5 μM EdU was added to culture medium, and samples were incubated for 18 hours. Samples were then fixed, permeabilized, and treated with Click-iT® EdU reaction cocktail according to kit instructions. Hoechst and/or Vybrant Dye (Life technologies) were diluted 1:1000 to measure DNA content.

FACs analysis was performed on a BD LSR II flow cytometry machine.

ChIP-seq Analysis

All FOXA2 ChIP-seq dataset from different conditions and cell-types were aligned using Bowtie 2 [222] to hg19 human genome reference assembly using default parameters. Duplicate reads were removed using Picard (http://broadinstitute.github.io/picard/). Genome browser images were created by converting bam files in .tdf files using IGV tools [223] by normalizing them to 1 million reads. All data sets were subjected to irreproducible discovery rate (IDR) framework [196] with 0.1 as cutoff in combination with using MACS2 [224] for calling peaks in each replicate separately. For MACS2 peak calling we used corresponding whole cell extract (WCE) as background control and p-value cutoff of 0.01. This initial peak calling using IDR and MACS2 resulted in a set of peaks that are above background for each cell-type. As an additional filtering and also for making peaks from different cell-types and conditions more easily comparable, we developed an in-house computational framework to redefine relative peak positions and

also peak equivalent peak width. IDR called peaks from all cell-types were merged together if they were found to be overlapping by at least 20% while keeping track of the summits of the peak that are being merged together. This resulted in a master peak set encompassing all FOXA2 datasets. As several peaks having different peak summits were merged together, we devised a simple weighted framework to define new peak submit. To assign a new peak summit we used the peak height as a measure of weighed distance from peak center. Using this weighted measure of peak height, we calculated a new peak summit which would be most close to the highest peak that was merged but will also represent contributions from smaller peaks in a distance dependent manner. All peaks were assigned new peak summits using this formula. To make the peak widths we transformed the coordinates using the new peak summit. We extended by 300bp in both directions from the peak summit to have all peaks of 600bp. Enrichment of different histone marks at these FOXA2 peaks was calculated using standard RPKM formula.

Composite plots

Composite plots to show enrichment of different histone marks at FOXA2 peaks were made by using Homer package [175] . As described in Homer documentation, we first created tag directories for each sample or histone mark we wish to plot around peak regions. Peaks were extended by 2000bp in each direction and tag directories were then used to create a matrix having tag densities at each nucleotide while normalizing each library for its respective library size. Matrix file having tag density at each position of extended 4000bp window was imported in R to create the plots.

Read Density Heatmaps

Read density heatmaps were created by using EnrichedHeatmap and ComplexHeatmap (https://github.com/jokergoo/EnrichedHeatmap, https://github.com/jokergoo/ComplexHeatmap) package. We first created genome-wide coverage of each sample or histone mark using coverageBed from BEDTools package [225]. These coverage files and the peaks regions that are required to be plotted were supplied as input to ComplexHeatmap. Heat on each heatmap was decided based on percentile range by capping the maximum at 99th percentile to remove outliers.

Differential Motif Analysis

Differential motif analysis was performed using Homer [175]. To calculate differential enrichment between two sets of peaks, we used provided one set as background and *vice versa*. The motifs were scanned in 200bp region around the peak center in both directions.

Chromatin State Maps

In order to classify FOXA2 bound regions in different chromatin states, we used a hierarchical classification system. All FOXA2 peaks that had either H3K27ac or H3K4me1 were marked as “active”. Excluding these “active” regions, regions having ATAC signal (RPKM) above 3 were marked as “open” and regions having H3K27me3 were classified as “repressed”. After classifying all histone marks, we divided rest of the regions based on their DNAme levels. Regions having DNAme levels below 20% were marked as low methylated regions (LMR), regions having methylation level between

20% and 60% were called intermediate methylated regions (IMR) while those above 60% methylation were termed as hypermethylated regions (HMR). The dynamics of transitions in chromatin state for each peak between different cell types (BJ, human ESC and dEN) was visualized using a heat map.

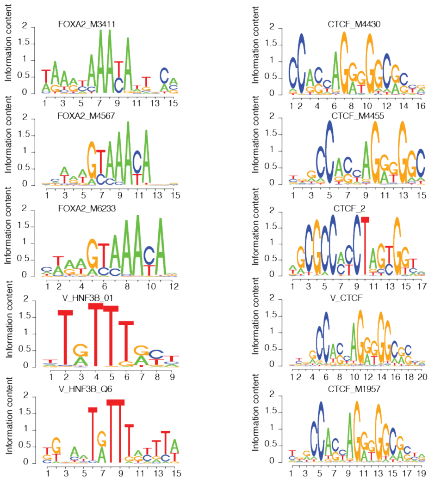
ChIP-BS-seq Analysis

For the analysis of methylation changes associated with FOXA2 binding, we redefined binding sites to maximize overlap with our ChIP-BS-seq data as bisulfite conversion on small amounts of input DNA results in degradation. To do this, we combined FOXA2 ChIP-seq data from BJ^{FOXA2} cells 4days and 10days post induction. The summit of each peak was determined using MACs, then the region 200bp either side was selected, and overlapping regions merged to generate a list of 113,398 sites. We then intersected these regions with BJ fibroblast WGBS and 4day BJ^{FOXA2} ChIP-BS-seq datasets and selected only CpGs covered by at least 3 reads in both samples, giving a total of 42,086 sites and 135,785 CpGs. We used these same regions to select matched CpGs covered at ≥3X in the ChIP-BS-seq data from the BJ^{FOXA2} mimosine treated and released samples (n = 13,494 sites and 18,429 CpGs). We compared individual CpG and mean FOXA2 binding site methylation, generated CpG count and coverage plots and calculated the distances between the summit and nearest CpG using custom R scripts. For comparison to ATAC-seq data (described above), we used HOMER [175] to generate enrichment composite plots for 2kb either side of our peaks.

Appendix

Parts of this chapter are submitted for publication elsewhere[1]

a



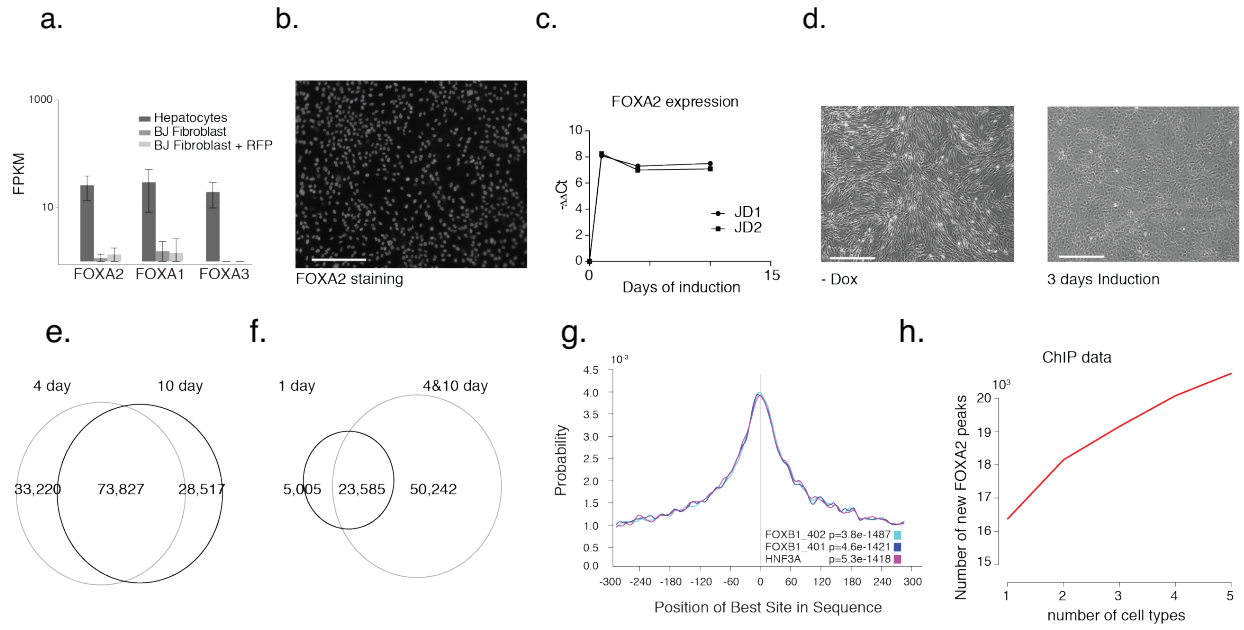
b

FOXA2	Total Motifs	Active Motifs	Active Bound Motifs	% Active Bound
M3411_1_02	1545443	324473	20367	6.277
M4567_1_02	1156736	336986	46035	13.661
M6233_1_02	1345871	329575	40108	12.170
V_HNF3B_01	1536836	322533	20268	6.284
V_HNF3B_Q6	1400606	360415	33423	9.273
POU5F1	Total Motifs	Active Motifs	Active Bound Motifs	% Active Bound
M1959_1_02	1419687	352888	17742	5.028
M6121_1_02	1072731	238237	10767	4.519
M6128_1_02	1289446	338978	16885	4.981
M6133_1_02	1711439	392192	17989	4.587
M4509_1_02	1253348	323039	17399	5.386
CTCF	Total Motifs	Active Motifs	Active Bound Motifs	% Active Bound
M4430_1_02	1871836	963097	92463	9.601
M4455_1_02	1672296	900740	88197	9.792
CTCF_2	1932729	850520	81390	9.569
V_CTCF_02	1036556	534044	68627	12.850
M1957_1_02	1866731	924066	88836	9.614

S1:

a) Motif logo's of the PWMs (shown in **Figure 1A**) used for identifying genome-wide occurrence of selected motifs throughout hg19 using FIMO [203].

b) Chart displaying name of PWM used in each motif analysis, number of times the PWM mapped across the genome, the number of motifs within potentially 'active' regulatory regions, motifs in 'active' regions bound by FOXA and the calculated percentage of bound motifs.



S2:

a) Expression bar plot displaying FPKM values for FOXA family members: FOXA2, FOXA1 and FOXA3 in human hepatocytes (positive control) BJ fibroblasts, BJ fibroblasts infected with control RFP virus (negative control). Error bars represent a 95% confidence interval around the average values.

b) Immunostaining for FOXA2 in the JD1 BJ^{FOXA2} line. 10X magnification shown. White scale bar is equal to 345nm.

c) qRT-PCR measurements of FOXA2 transcript level at four time points over a 10 day time course. No expression is measured on day 0. Stable FOXA2 transcript level is seen across days 1, 4 and 10 following induction.

d) Bright-field images show morphological change in JD1 BJ^{FOXA2} cells after 3 days of doxycycline. White scale bar is equal to 345nm.

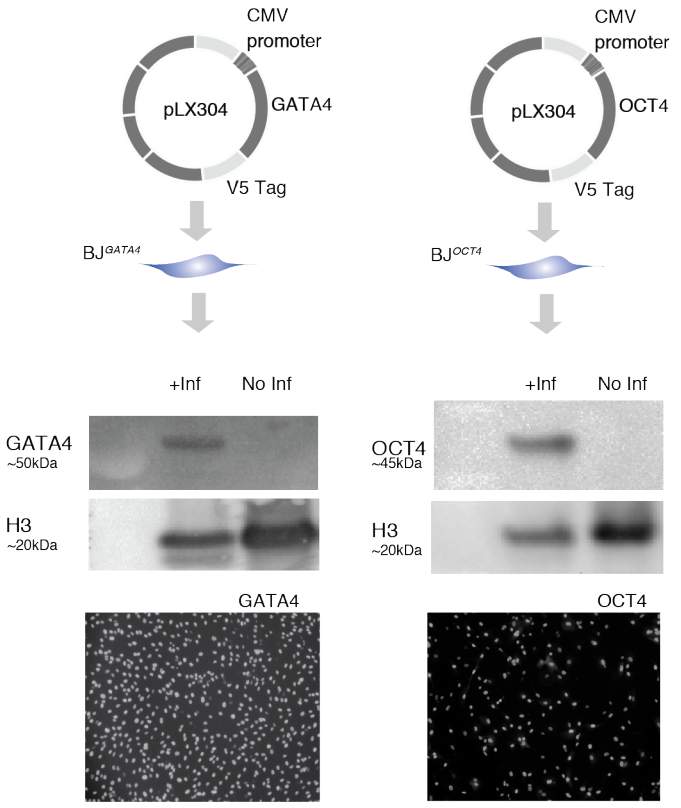
e) Venn diagram displays the strong overlap and similar number of MACS peak calls for FOXA2 ChIP-sequencing after 4 and 10 days of FOXA2 induction.

f) Venn diagram demonstrating the overlap of the intersection in MACS peak calls between the BJ^{FOXA2} day 4/day 10 time points (combined n=73,827) and FOXA2 ChIP-sequencing after 1 day of FOXA2 induction.

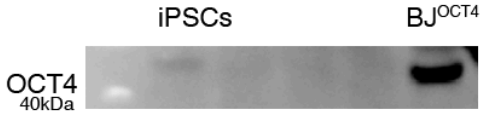
g) Centrimo output display of motif enrichment analysis at ectopic FOXA2 binding sites

h) Saturation analysis showing continuous gain in binding sites with FOXA2 ChIP-seq data from new cell types

a.

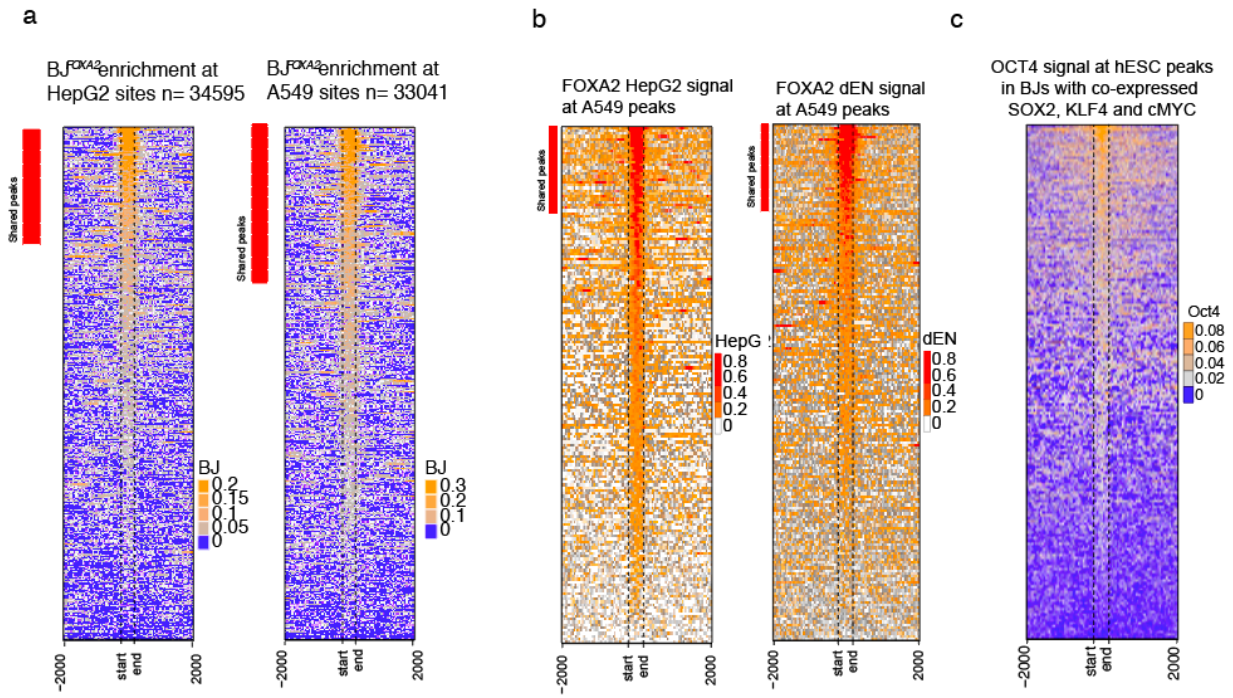


b.



S3:

a) Schematic of OCT4 and GATA4 ectopic systems with corresponding western blots demonstrating protein levels.
b) Western blot of ectopic and endogenous OCT4 protein levels in iPSCs and 4 days of dox BJ^{OCT4}



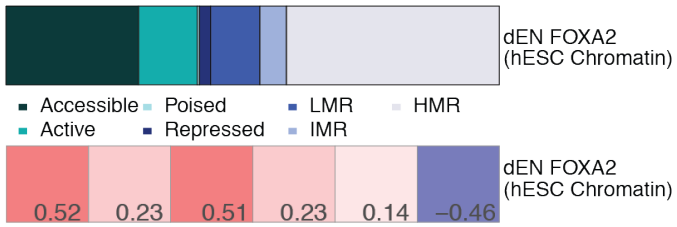
S4:

a) Read density heat maps of FOXA2 enrichment in BJ^{FOXA2} ChIP-sequencing data at FOXA2 endogenous peak called regions from HepG2 (n=34,595) and A549 (n= 33,041) cells. Bars indicates peak calls in common between ectopic FOXA2 ChIP-sequencing data and endogenous (HepG2 or A549) FOXA2 ChIP-sequencing. Dashed lines represent the start and end of FOXA2 peaks with 2kb extension on either side of the peak. Similar to the dEN results, most HepG2 and A549 sites still show some level of enrichment of FOXA2 in BJs that are however not called as significantly enriched by our MACS peak calling.

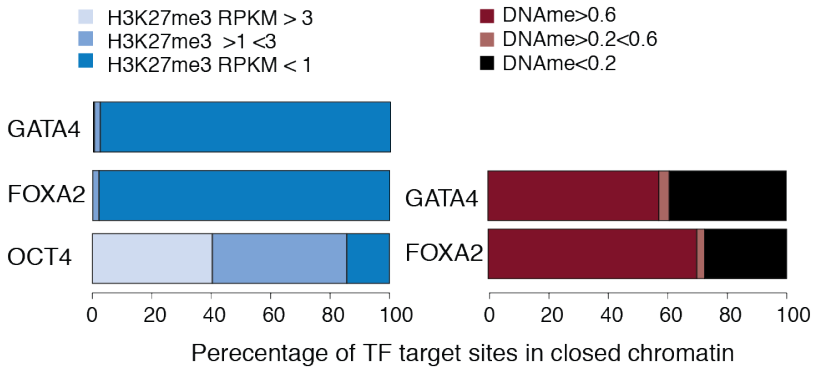
b) Endogenous sampling demonstrated by read density heat maps if FOXA2 enrichment in HepG2 and dEN at A549 bound FOXA2 sites. Bar indicates peak calls in common between HepG2 and dEN FOXA2 ChIP-sequencing data and A549 FOXA2 ChIP-sequencing. Dashed lines represent the start and end of FOXA2 peaks with 2kb extension on either side of the peak.

c) Read density heat maps of OCT4 enrichment in BJ cells infected with OCT4, SOX2, KLF4 and cMYC [21] at OCT4 bound regions in human ESCs. Dashed lines represent the start and end of OCT4 peaks with 2kb extension on either side of the peak.

a.



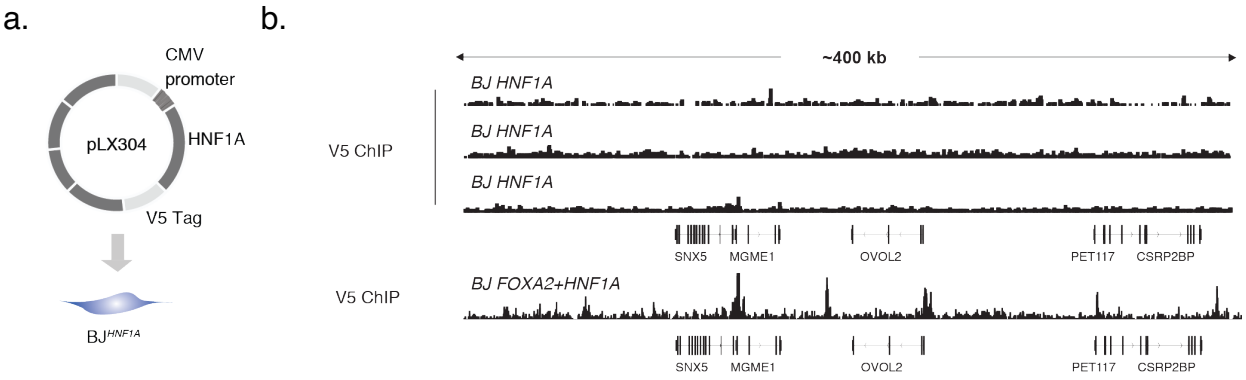
b.



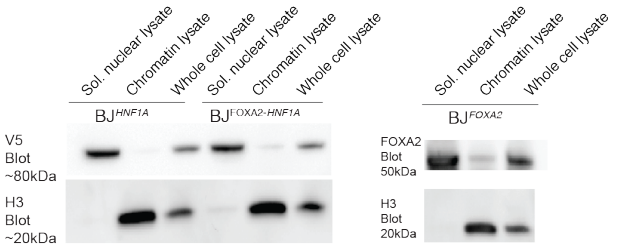
S5:

a) Chromatin state map defining percentages of dEN FOXA2 bound regions using human ESC chromatin data. Spearman correlation with dEN FOXA2 peaks and human ESC chromatin.

b) Left: Stacked bar plots display FOXA2, GATA4, OCT4 closed chromatin bound regions and levels of H3K27me3. Right: Stacked bar plot displays levels of DNAme at FOXA2 and GATA4 bound regions in closed chromatin.

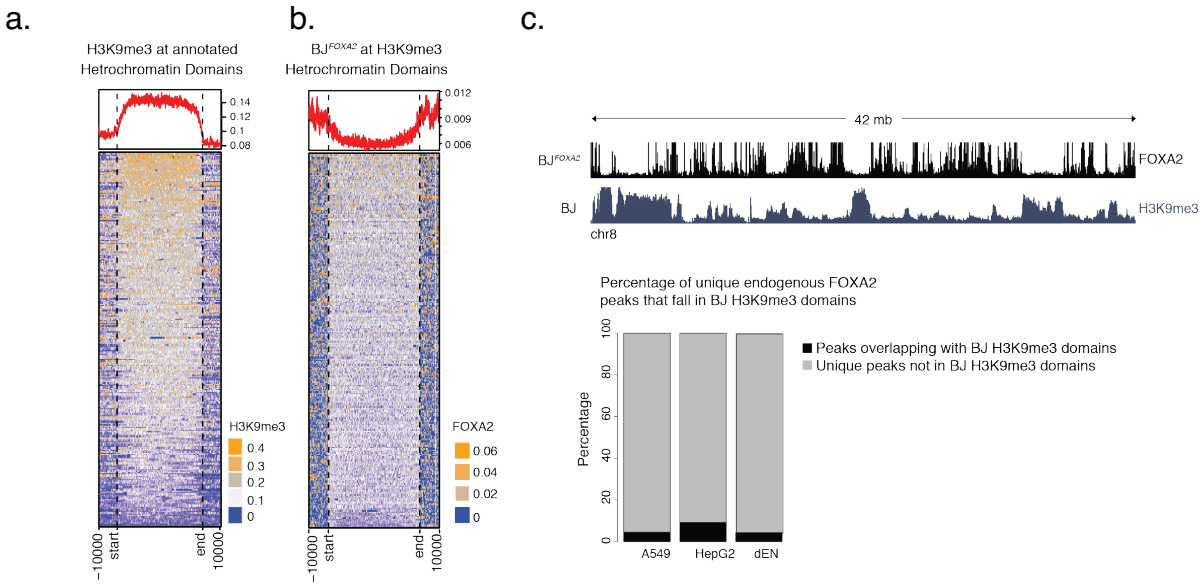


c.



S6:

a) Schematic representation of system used to generate BJ^{HNF1A} cells.
b) IGV browser shots displaying a 400 kB genomic region in HNF1A (using V5 antibody) ChIP experiments. Top three experiments are distinct biological replicate experiments in BJ^{HNF1A} cells. In contrast, the bottom track represents HNF1A binding when FOXA2 is co-expressed.
c) Western blot analysis of HNF1A (v5) protein levels in soluble nuclear, chromatin bound and whole cell lysates in BJ^{HNF1A} cells compared to BJ$^{FOXA2\text{-}HNF1A}$ cells. Control blots in BJ^{FOXA2} cells demonstrate distinct difference in chromatin bound protein fraction of the two factors assessed.

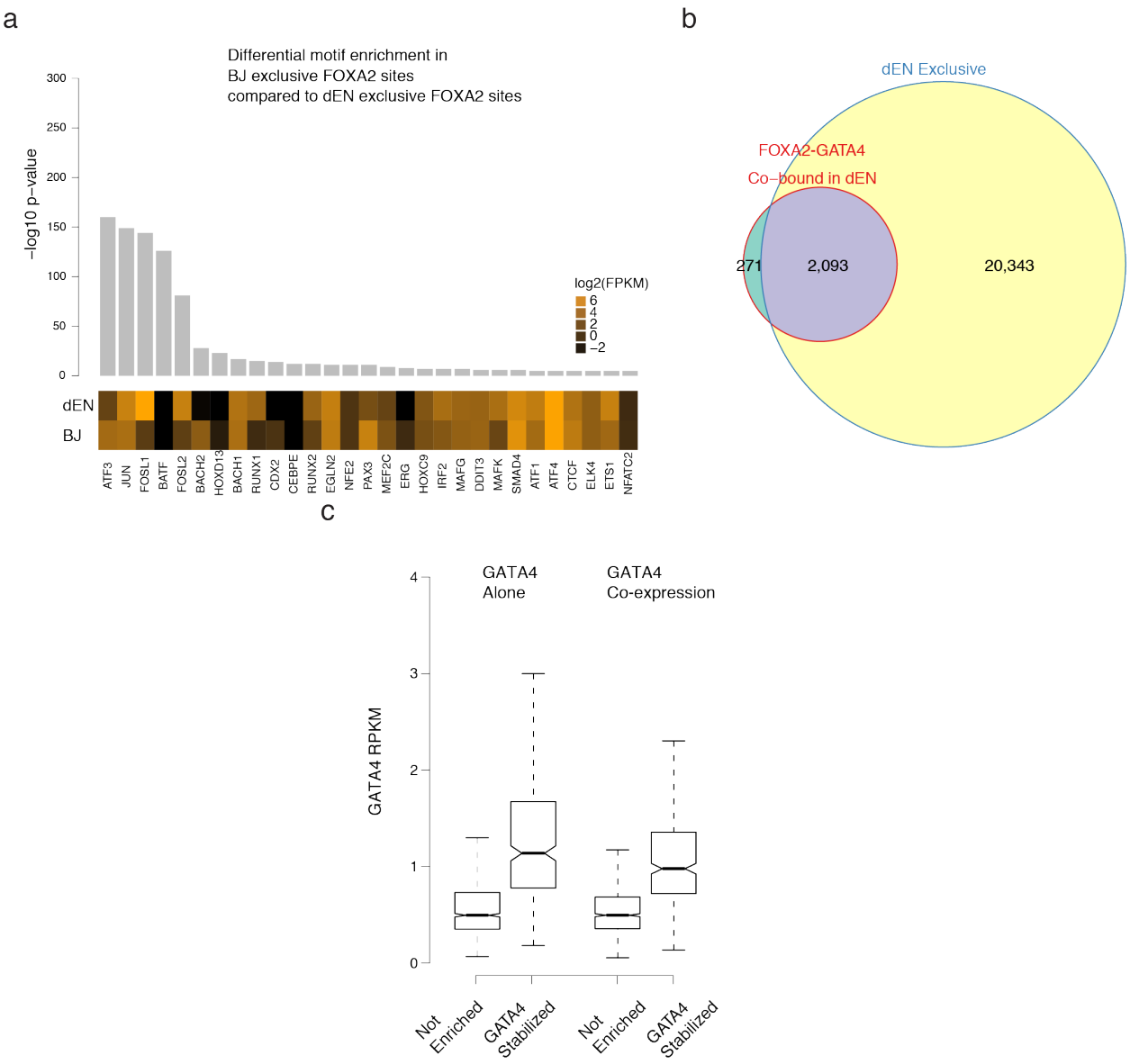


S7:

a) Read density heat map displaying enrichment of BJ H3K9me3 ChIP-sequencing (REMC) at heterochromatin domains (n=256) defined in [21].

b) Read density heat map displaying FOXA2 enrichment of BJ FOXA2 ChIP-sequencing (REMC) at heterochromatin domains defined in [21].

c) Representative IGV browser tracks showing a zoomed out view on chromosome 8 (305,736- 42,374,902) that visualizes the general depletion of FOXA2 binding within H3K9me3 marked regions. Below displays the percentage of exclusively bound endogenous sites that are found in K9-domains.

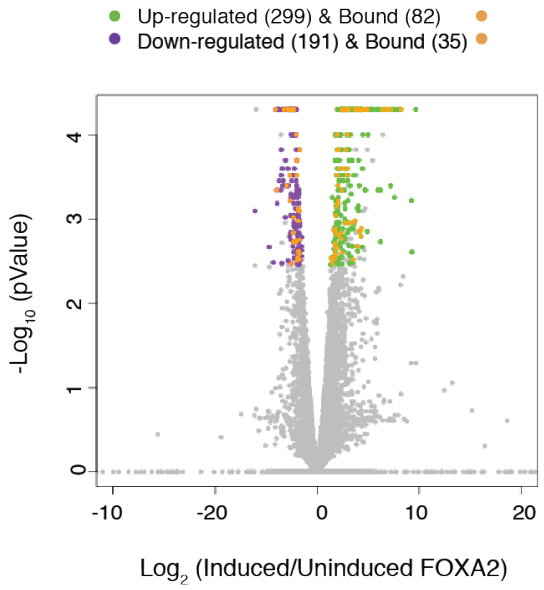


S8:

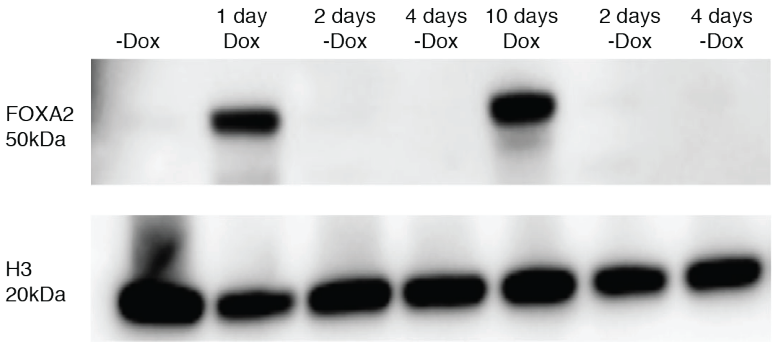
a) Differential motif analysis displaying $-\log_{10}$ *p*-value of enriched motifs in BJ exclusive sites versus dEN exclusive sites with the most significant motifs on the left. Expression ($\log_2$ FPKM) of the TF associated with the motif in both BJ and dEN is shown on the bottom.

b) Venn diagram showing the overlap between IDR peak calls that are co-bound by FOXA2 and GATA4 in dEN and dEN exclusive targets that are not occupied in BJs.

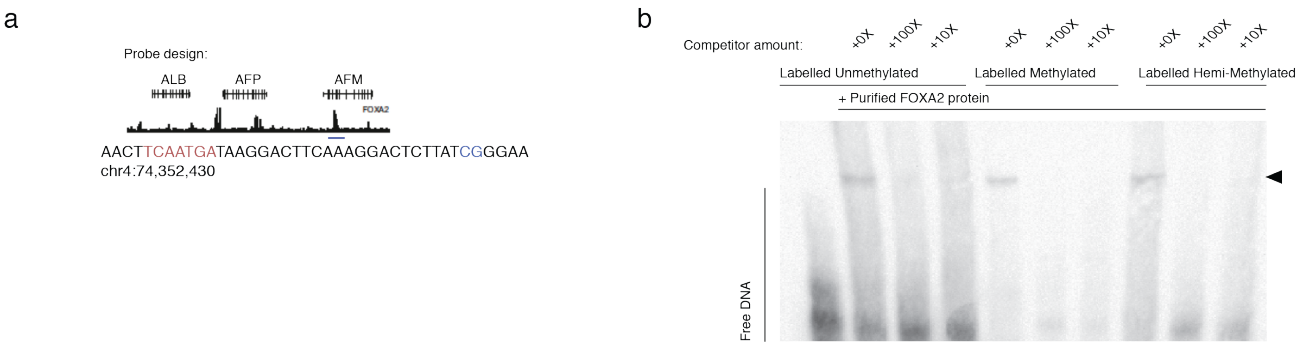
c) Bar plots displaying the RPKM of GATA4 enrichment in BJ^{GATA4} and $BJ^{FOXA2\text{-}GATA4}$ at the subset of regions that are GATA4 stabilized compared to the non-enriched subset. Boxes indicate interquartile range and whiskers show maximum and minimum values. Outliers are removed.



S9:
Volcano plot of differentially expressed genes on day 4 of FOXA2 induction in the BJ^{FOXA2} line compared to the uninduced control. Differentially expressed genes are identified using cufflinks. Y-axis represents $-\log_{10}$ of *p*-value while x-axis shows fold change in $\log_2$ scale.



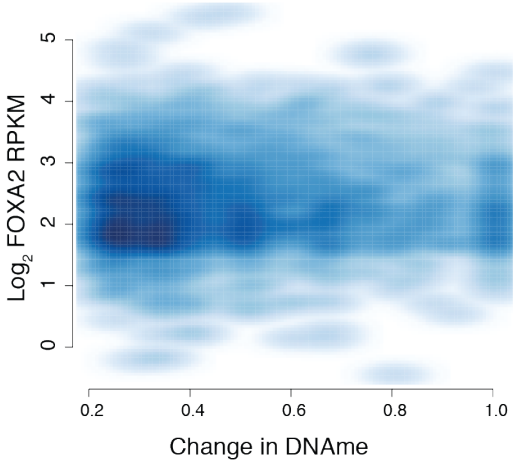
S10:
Western blot analysis of FOXA2 and H3 protein as loading control after FOXA2 induction for 1 day and 10 days followed by 2 and 4 days of doxycycline withdrawal.



S11:

a) Oligomer probes were designed for ElectroMobility Shift Assay (EMSA) at the FOXA2 binding sites in the AFM genes as shown in the IGV browser track (chr4:74,263,092-74,395,230). Two oligo versions were synthesized for AFM - with and without a methylated CpG. Motif sequence is highlighted in Red and CpG in Blue.

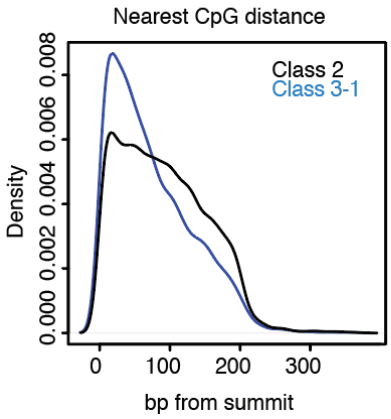
b) EMSA using purified Halo-tagged FOXA2 protein demonstrates FOXA2 interacts equally with methylated, hemi-methylated and non-methylated oligomers. Competition experiments were performed with non-biotinlyated oligomers at 10X and 100X the concentration of the biotinlyated oligomers.



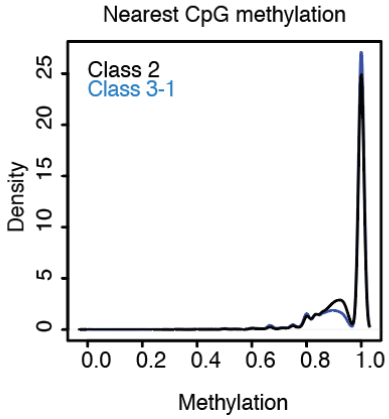
S12:

Scatter plot of FOXA2 enrichment at class 3-1 regions compared to their change in DNAme

a



b

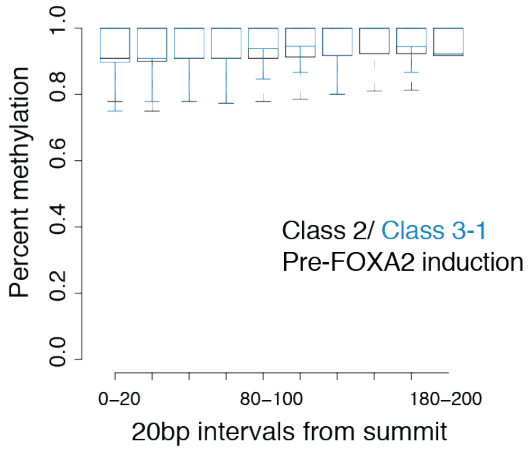


c

	Distance to nearest CpG	% methylation nearest CpG
Class 2	90.27212**	.949612
Class 3-1	73.70712**	.949471

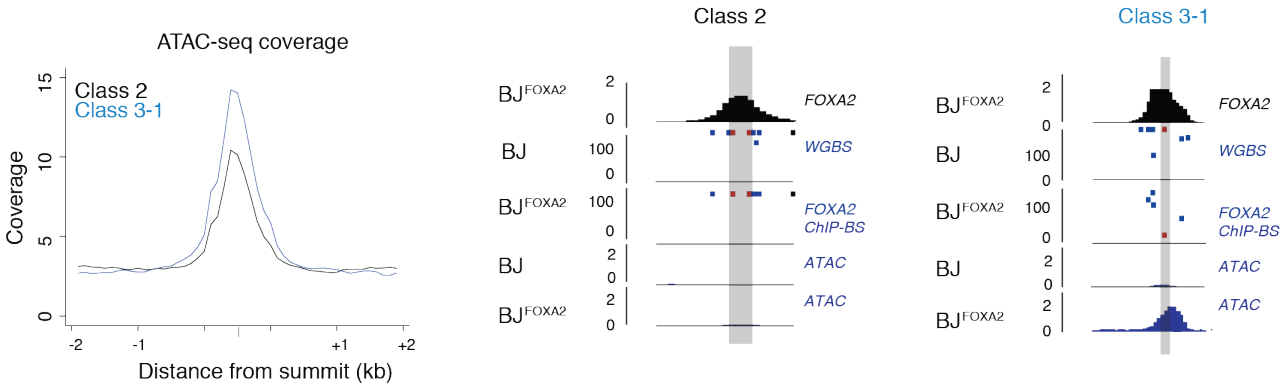
S13:

a) Density plot capturing distance to nearest CpG from the summit of FOXA2 ChIP-sequencing peaks. Class 2 black. Class 3-1 blue.
b) Density plot capturing the percent methylation of the nearest CpG from the summit of FOXA2 ChIP-sequencing peaks. Class 2 black. Class 3-1 blue.
c) Average distance and methylation status to the nearest CpG from the peak summit. Statistical significance shown by Welch *t*-test.

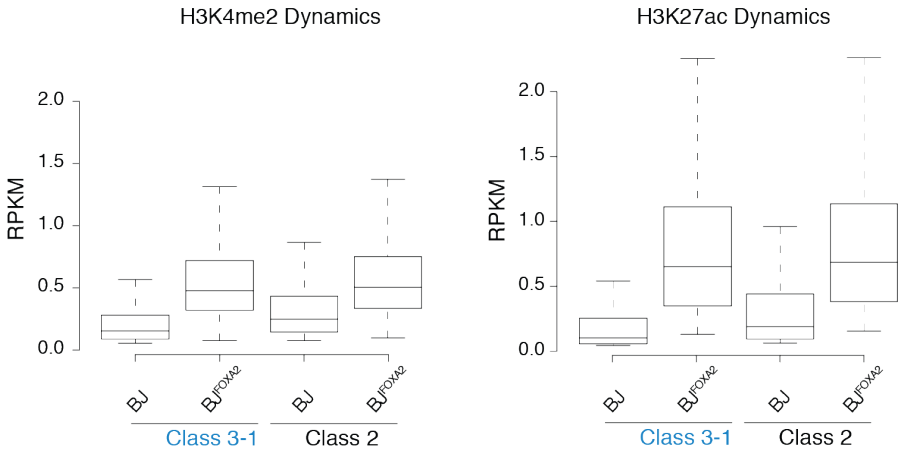


S14:

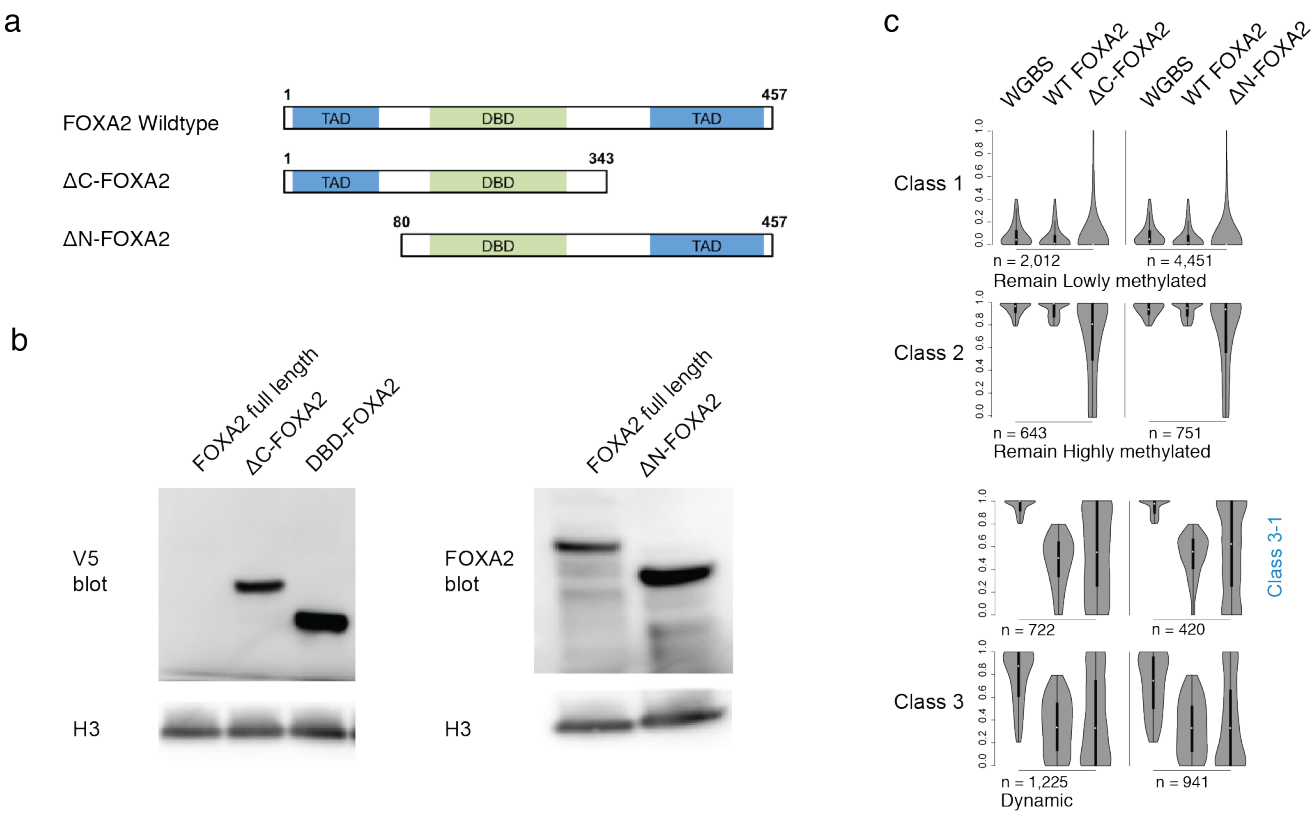
Box plot shows the percent methylation of CpGs within 20bp windows from the summit of the peak extended 200bp. Methylation measurements were taken from WGBS data prior to FOXA2 induction. Class 2 black. Class 3-1 blue. Boxes indicate interquartile range and whiskers show maximum and minimum values. Outliers are removed.



S15:
Density plot of ATAC-seq coverage 2 days after FOXA2 induction for Class 2 (black) and Class 3-1 (blue) target sites (left). Accompanying browser tracks on the right display FOXA2 ChIP-sequencing, BJ WGBS, FOXA2 ChIP-BS, and ATAC-seq prior to FOXA2 induction as well as 2 days following the induction. Class 2 region shown is chr12:54,011,044-54,012,658 and Class 3-1 region shown is chr1:28,720,983-28,722,960. CpGs included in the analysis are shown in red and highlighted by a gray box.

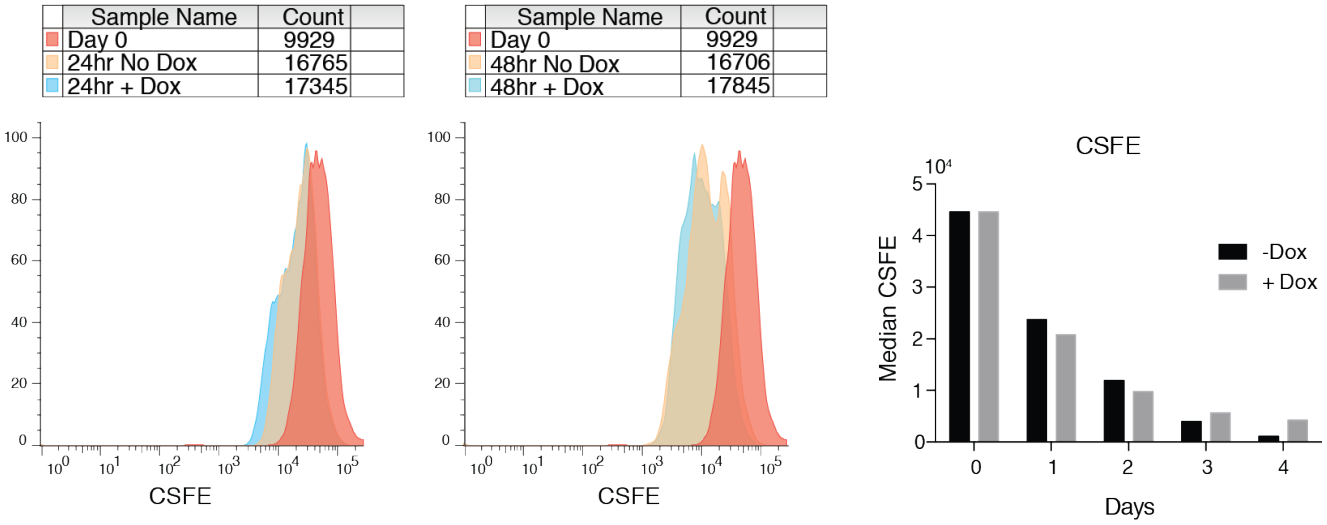


S16:
Box plots displaying mean RPKM values of H3K4me2 (left) and H3K27ac (right) at class 3-1 compared to class 2 targets in pre- and post- FOXA2 induction conditions.

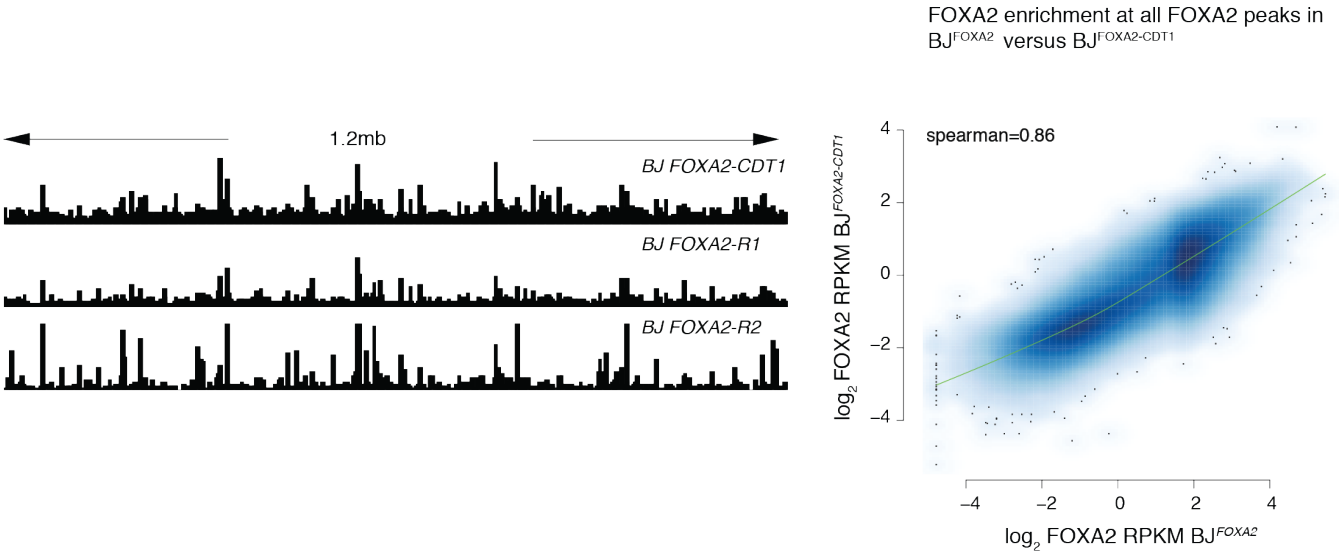


S17:

a) Schematic representation of V5-tagged FOXA2 deletion constructs
b) Western blots displaying protein levels of FOXA2 full length and deletions construct proteins with V5 (left) and FOXA2 (right) blots with H3 loading control. deltaN construct was better detected with FOXA2 antibody.
c) Violin plots of DNAme levels following FOXA2 ChIP-BS-seq of deletion constructs at class 1, 2 and 3 targets.

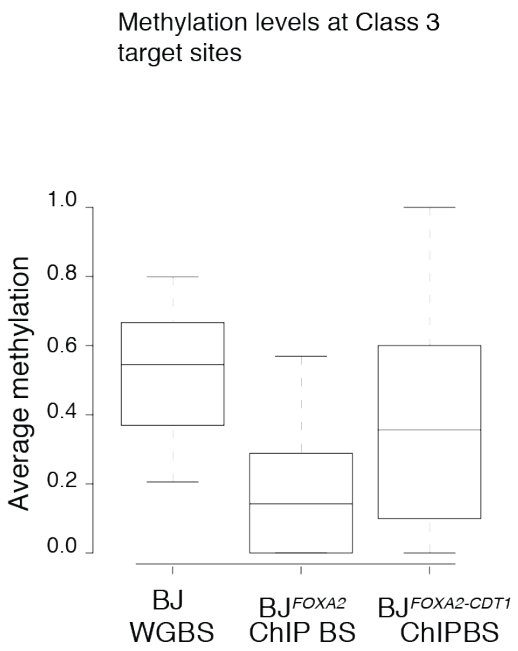


S18:
CSFE time course signal for samples after 24 hours and 48 hours labeling plus/minus FOXA2 induction overlaid on Day 0 labeling time point. Bar plot shows the median CSFE signal for cells plus/minus dox induction of FOXA2 over 4 days.



S19:

a) IGV browser shot of a 1.2mB genomic region (chr1:92,780,270-94,073,813) in BJ$^{FOXA2\text{-}CDT1}$ compared to two replicates of BJ^{FOXA2} FOXA2 ChIP-seq.
b) Scatter plot displaying FOXA2 enrichment value at union FOXA2 peak set for BJ$^{FOXA2\text{-}CDT1}$ compared to BJ^{FOXA2} FOXA2 ChIP-seq.



S20:
Box plots show average methylation of all Class 3 in BJ WGBS, BJ^{FOXA2} ChIP-BS and BJ$^{FOXA2\text{-}CDT1}$ ChIP-BS data. Regions shown had at least 10X coverage. Boxes indicate interquartile range and whiskers show maximum and minimum values. Outliers are removed.

References

1. Julie Donaghey, S.T., Jocelyn Charlton, Jennifer Chen, Zachary D. Smith, Hongcang Gu, Ramona Pop, Kendell Clement, Elena Stamenova, David R. Kelley, Casey A. Gifford, Davide Cacchiarelli, John L. Rinn, Andreas Gnirke, Michael J. Ziller, Alexander Meissner. Genetic determinants and epigenetic effects of pioneer factor occupancy. *Nature Genetics* **under review** (2017).

2. Zaret, K.S. & Carroll, J.S. Pioneer transcription factors: establishing competence for gene expression. *Genes & development* **25**, 2227-2241 (2011).

3. Iwafuchi-Doi, M. *et al.* The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. *Molecular cell* **62**, 79-91 (2016).

4. Iwafuchi-Doi, M. & Zaret, K.S. Pioneer transcription factors in cell reprogramming. *Genes & development* **28**, 2679-2692 (2014).

5. Spitz, F. & Furlong, E.E. Transcription factors: from enhancer binding to developmental control. *Nature reviews. Genetics* **13**, 613-626 (2012).

6. Cirillo, L.A. *et al.* Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome. *The EMBO journal* **17**, 244-254 (1998).

7. McPherson, C.E., Horowitz, R., Woodcock, C.L., Jiang, C. & Zaret, K.S. Nucleosome positioning properties of the albumin transcriptional enhancer. *Nucleic acids research* **24**, 397-404 (1996).

8. Cirillo, L.A. *et al.* Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Molecular cell* **9**, 279-289 (2002).

9. Gualdi, R. *et al.* Hepatic specification of the gut endoderm in vitro: cell signaling and transcriptional control. *Genes & development* **10**, 1670-1682 (1996).

10. Ang, S.L. *et al.* The formation and maintenance of the definitive endoderm lineage in the mouse: involvement of HNF3/forkhead proteins. *Development (Cambridge, England)* **119**, 1301-1315 (1993).

11. Ang, S.L. & Rossant, J. HNF-3 beta is essential for node and notochord formation in mouse development. *Cell* **78**, 561-574 (1994).

12. Gao, N. *et al.* Foxa1 and Foxa2 maintain the metabolic and secretory features of the mature beta-cell. *Molecular endocrinology (Baltimore, Md.)* **24**, 1594-1604 (2010).

13. Wang, A. *et al.* Epigenetic priming of enhancers predicts developmental competence of human ESC-derived endodermal lineage intermediates. *Cell stem cell* **16**, 386-399 (2015).

14. Lupien, M. *et al.* FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132**, 958-970 (2008).

15. Zaret, K.S. & Mango, S.E. Pioneer transcription factors, chromatin dynamics, and cell fate control. *Current opinion in genetics & development* **37**, 76-81 (2016).

16. 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. *Nature genetics* **43**, 27-33 (2011).

17. Chen, J. *et al.* Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. *Cell* **156**, 1274-1285 (2014).

18. Swinstead, E.E. *et al.* Steroid Receptors Reprogram FoxA1 Occupancy through Dynamic Chromatin Transitions. *Cell* **165**, 593-605 (2016).

19. Liu, Z. & Kraus, W.L. Catalytic-Independent Functions of PARP-1 Determine Sox2 Pioneer Activity at Intractable Genomic Loci. *Molecular cell* **65**, 589-1704578560 (2017).

20. Franco, H.L., Nagari, A. & Kraus, W.L. TNFα signaling exposes latent estrogen receptor binding sites to alter the breast cancer cell transcriptome. *Molecular cell* (2015).

21. 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).

22. Soufi, A. *et al.* Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161**, 555-568 (2015).

23. 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. *The Journal of biological chemistry* **285**, 16135-16144 (2010).

24. Sérandour, A.A. *et al.* Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers. *Genome research* **21**, 555-565 (2011).

25. Gifford, C.A. *et al.* Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell* **153**, 1149-1163 (2013).

26. Koch, F. *et al.* Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nature Structural & Molecular Biology* **18**, 956-963 (2011).

27. Szutorisz, H., Dillon, N. & Tora, L. The role of enhancers as centres for general transcription factor recruitment. *Trends in Biochemical Sciences* **30**, 593-599 (2005).

28. Pennacchio, L.A., Bickmore, W., Dean, A., Nobrega, M.A. & Bejerano, G. Enhancers: five essential questions. *Nature Reviews Genetics* **14**, 288-295 (2013).

29. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

30. Yan, J. *et al.* Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* **154**, 801-813 (2013).

31. Boyer, L.A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947-956 (2005).

32. Thanos, D. & Maniatis, T. Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell* **83**, 1091-1100 (1995).

33. Pan, Y., Tsai, C.-J.J., Ma, B. & Nussinov, R. Mechanisms of transcription factor selectivity. *Trends in genetics : TIG* **26**, 75-83 (2010).

34. Yang, A. *et al.* Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells. *Molecular cell* **24**, 593-602 (2006).

35. Joseph, R. *et al.* Integrative model of genomic factors for determining binding site selection by estrogen receptor-α. *Molecular systems biology* **6**, 456 (2010).

36. Farnham, P.J. Insights from genomic profiling of transcription factors. *Nature Reviews Genetics* **10**, 605-616 (2009).

37. Dror, I., Golan, T., Levy, C., Rohs, R. & Mandel-Gutfreund, Y. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome research* **25**, 1268-1280 (2015).

38. Slattery, M. *et al.* Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* **147**, 1270-1282 (2011).

39. Slattery, M. *et al.* Absence of a simple code: how transcription factors read the genome. *Trends in biochemical sciences* **39**, 381-399 (2014).

40. Stormo, G.D. & Zhao, Y. Determining the specificity of protein–DNA interactions. *Nature Reviews Genetics* (2010).

41. von Hippel, P.H. & Berg, O.G. Facilitated target location in biological systems. *The Journal of biological chemistry* **264**, 675-678 (1989).

42. Jolma, A. *et al.* DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384-388 (2015).

43. Bulyk, M.L. Computational prediction of transcription-factor binding site locations. *Genome biology* **5**, 201 (2003).

44. Berger, M.F., Philippakis, A.A., Qureshi, A.M. & He, F.S. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature …* (2006).

45. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327-339 (2013).

46. Gordân, R. *et al.* Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell reports* **3**, 1093-1104 (2013).

47. Dror, I., Zhou, T., Mandel-Gutfreund, Y. & Rohs, R. Covariation between homeodomain transcription factors and the shape of their DNA binding sites. *Nucleic acids research* **42**, 430-441 (2014).

48. Luger, K., Mäder, A., Richmond, R., Sargent, D. & Richmond, T. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251-260 (1997).

49. Allan, J., Hartman, P.G., Crane-Robinson, C. & Aviles, F.X. The structure of histone H1 and its location in chromatin. *Nature* **288**, 675-679 (1980).

50. Kelly, T.K. *et al.* Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome research* **22**, 2497-2506 (2012).

51. He, H. *et al.* Nucleosome dynamics define transcriptional enhancers. *Nature Genetics* **42**, 343-347 (2010).

52. Henikoff, S. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nature reviews. Genetics* **9**, 15-26 (2008).

53. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772-778 (2006).

54. Struhl, K. & Segal, E. Determinants of nucleosome positioning. *Nature structural & molecular biology* **20**, 267-273 (2013).

55. Drew, H.R. & Travers, A.A. DNA bending and its relation to nucleosome positioning. *Journal of molecular biology* **186**, 773-790 (1985).

56. Nelson, H.C., Finch, J.T., Luisi, B.F. & Klug, A. The structure of an oligo(dA).oligo(dT) tract and its biological implications. *Nature* **330**, 221-226 (1987).

57. Iyer, V. & Struhl, K. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *The EMBO journal* **14**, 2570-2579 (1995).

58. Zhang, Z. *et al.* A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science (New York, N.Y.)* **332**, 977-980 (2011).

59. Zhang, Y. *et al.* Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nature structural & molecular biology* **16**, 847-852 (2009).

60. Weber, C.M. & Henikoff, S. Histone variants: dynamic punctuation in transcription. *Genes & development* **28**, 672-682 (2014).

61. Suto, R.K., Clarkson, M.J., Tremethick, D.J. & Luger, K. Crystal structure of a nucleosome core particle containing the variant histone H2A.Z. *Nature structural biology* **7**, 1121-1124 (2000).

62. Zlatanova, J. & Thakar, A. H2A.Z: view from the top. *Structure (London, England : 1993)* **16**, 166-179 (2008).

63. Goldman, J.A., Garlick, J.D. & Kingston, R.E. Chromatin remodeling by imitation switch (ISWI) class ATP-dependent remodelers is stimulated by histone variant H2A. Z. *Journal of Biological Chemistry* **285**, 4645-4651 (2010).

64. Ahmad, K. & Henikoff, S. The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. *Molecular cell* **9**, 1191-1200 (2002).

65. Mito, Y., Henikoff, J.G. & Henikoff, S. Histone replacement marks the boundaries of cis-regulatory domains. *Science* (2007).

66. Schwartz, B.E. & Ahmad, K. Transcriptional activation triggers deposition and removal of the histone variant H3. 3. *Genes & development* (2005).

67. Grunstein, M. Histone acetylation in chromatin structure and transcription. *Nature* **389**, 349-352 (1997).

68. Shogren-Knaak, M. *et al.* Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science (New York, N.Y.)* **311**, 844-847 (2006).

69. Azuara, V. *et al.* Chromatin signatures of pluripotent cell lines. *Nature cell biology* **8**, 532-538 (2006).

70. Bock, C. *et al.* DNA Methylation Dynamics during In Vivo Differentiation of Blood and Skin Stem Cells. *Molecular cell* **47**, 633-647 (2012).

71. Consortium, E.P. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).

72. Dixon, J. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012).

73. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49 (2011).

74. Hawkins, R. *et al.* Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell stem cell* **6**, 479-491 (2010).

75. Heintzman, N. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-112 (2009).

76. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**, 576-589 (2010).

77. Hon, G., Wang, W. & Ren, B. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS computational biology* **5**(2009).

78. Casey A. Gifford, M.J.Z., Alexander Tsankov, Honcang Gu, Robbyn Issner, Xiaolan Zhang, Michael Coyne, Jennifer L. Fostel, Laurie Holmes, Jim Meldrim, Charles Epstein, Oliver Kohlbacher, Andreas Gnirke, Bradley E. Bernstein, and Alexander Meissner Epigenetic and Transcriptional Dynamics During Lineage Specification of Human Embryonic Stem Cells *Submitted to Cell* (2012).

79. Creyghton, M.P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-6 (2010).

80. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-9 (2011).

81. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-12 (2009).

82. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279-83 (2011).

83. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279-283 (2011).

84. Thurman, R. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).

85. Creyghton, M. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 21931-21936 (2010).

86. Mikkelsen, T. *et al.* Comparative epigenomic analysis of murine and human adipogenesis. *Cell* **143**, 156-169 (2010).

87. Ram, O. *et al.* Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. *Cell* **147**, 1628-1639 (2011).

88. Taberlay, P. *et al.* Polycomb-repressed genes have permissive enhancers that initiate reprogramming. *Cell* **147**, 1283-1294 (2011).

89. Mikkelsen, T. *et al.* Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454**, 49-55 (2008).

90. Mieczkowski, J. *et al.* MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nature Communications* **7**, 11485 (2016).

91. Mueller, B. *et al.* Widespread changes in nucleosome accessibility without changes in nucleosome occupancy during a rapid transcriptional induction. *Genes & Development* **31**, 451-462 (2017).

92. Smith, Z.D. & Meissner, A. DNA methylation: roles in mammalian development. *Nature reviews. Genetics* **14**, 204-220 (2013).

93. Ziller, M.J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477-481 (2013).

94. Jones, P.A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics* **13**, 484-492 (2012).

95. Stadler, M. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490-495 (2011).

96. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-322 (2009).

97. Koche, R. *et al.* Reprogramming factor expression initiates widespread targeted chromatin remodeling. *Cell stem cell* **8**, 96-201 (2011).

98. Schübeler, D. Function and information content of DNA methylation. *Nature* **517**, 321-326 (2015).

99. Domcke, S. *et al.* Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528**, 575-579 (2015).

100. Medvedeva, Y.A. *et al.* Effects of cytosine methylation on transcription factor binding sites. *BMC Genomics* **15**, 1-12 (2013).

101. Stadler, M.B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490-495 (2011).

102. Wang, H. *et al.* Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome research* **22**, 1680-1688 (2012).

103. Flavahan, W.A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110-114 (2016).

104. Maurano, M.T. *et al.* Role of DNA Methylation in Modulating Transcription Factor Occupancy. *Cell reports* **12**, 1184-1195 (2015).

105. Hu, S. *et al.* DNA methylation presents distinct binding sites for human transcription factors. *eLife* **2**(2013).

106. Zhu, H., Wang, G. & Qian, J. Transcription factors as readers and effectors of DNA methylation. *Nature reviews. Genetics* **17**, 551-565 (2016).

107. Gifford, C. & Meissner, A. Epigenetic obstacles encountered by transcription factors: reprogramming against all odds. *Current opinion in genetics & development* (2012).

108. You, J. *et al.* OCT4 establishes and maintains nucleosome-depleted regions that provide additional layers of epigenetic regulation of its target genes. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 14497-14502 (2011).

109. Smith, Z.D. & Meissner, A. The simplest explanation: passive DNA demethylation in PGCs. *The EMBO journal* **32**, 318-321 (2013).

110. Saitou, M., Barton, S.C. & Surani, M.A. A molecular programme for the specification of germ cell fate in mice. *Nature* **418**, 293-300 (2002).

111. Lees-Murdock, D.J., De Felici, M. & Walsh, C.P. Methylation dynamics of repetitive DNA elements in the mouse germ cell lineage. *Genomics* **82**, 230-237 (2003).

112. Feng, S., Jacobsen, S.E. & Reik, W. Epigenetic Reprogramming in Plant and Animal Development. *Science* **330**, 622-627 (2010).

113. Kagiwada, S., Kurimoto, K., Hirota, T., Yamaji, M. & Saitou, M. Replication-coupled passive DNA demethylation for the erasure of genome imprints in mice. *The EMBO journal* **32**, 340-353 (2013).

114. Tang, W.W., Kobayashi, T., Irie, N., Dietmann, S. & Surani, M.A. Specification and epigenetic programming of the human germ line. *Nature reviews. Genetics* **17**, 585-600 (2016).

115. Kohli, R.M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502**, 472-479 (2013).

116. Wu, H. & Zhang, Y. Reversing DNA methylation: mechanisms, genomics, and biological functions. *Cell* **156**, 45-68 (2014).

117. Wu, S.C. & Zhang, Y. Active DNA demethylation: many roads lead to Rome. *Nature reviews. Molecular cell biology* **11**, 607-620 (2010).

118. Gehring, M., Reik, W. & Henikoff, S. DNA demethylation by DNA repair. *Trends in genetics : TIG* **25**, 82-90 (2009).

119. Cortázar, D. *et al.* Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability. *Nature* **470**, 419-423 (2011).

120. Cortellino, S. *et al.* Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell* **146**, 67-79 (2011).

121. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930-935 (2009).

122. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science (New York, N.Y.)* **333**, 1300-1303 (2011).

123. He, Y.-F.F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science (New York, N.Y.)* **333**, 1303-1307 (2011).

124. Globisch, D. *et al.* Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PloS one* **5**(2010).

125. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* (2009).

126. Valinluck, V. & Sowers, L.C. Endogenous cytosine damage products alter the site selectivity of human DNA maintenance methyltransferase DNMT1. *Cancer research* **67**, 946-950 (2007).

127. Hashimoto, H. *et al.* Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic acids research* **40**, 4841-4849 (2012).

128. Rasmussen, K.D. & Helin, K. Role of TET enzymes in DNA methylation, development, and cancer. *Genes & development* **30**, 733-750 (2016).

129. Sérandour, A.A.A. *et al.* Dynamic hydroxymethylation of deoxyribonucleic acid marks differentiation-associated enhancers. *Nucleic acids research* **40**, 8255-8265 (2012).

130. Ficz, G. *et al.* Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398-402 (2011).

131. Pastor, W.A. *et al.* Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473**, 394-397 (2011).

132. Cirillo, L. *et al.* Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. *Molecular cell* **9**, 279-289 (2002).

133. Cirillo, L. *et al.* Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome. *The EMBO journal* **17**, 244-254 (1998).

134. Cirillo, L. & Zaret, K. An early developmental transcription factor complex that is more stable on nucleosome core particles than on free DNA. *Molecular cell* **4**, 961-969 (1999).

135. Bossard, P. & Zaret, K.S. GATA transcription factors as potentiators of gut endoderm differentiation. *Development (Cambridge, England)* **125**, 4909-4917 (1998).

136. Clark, K., Halay, E., Lai, E. & Burley, S. Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. *Nature* **364**, 412-420 (1993).

137. 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**, 412-420 (1993).

138. Allan, J., Hartman, P., Crane-Robinson, C. & Aviles, F. The structure of histone H1 and its location in chromatin. *Nature* **288**, 675-679 (1980).

139. Monaghan, A., Kaestner, K., Grau, E. & Schütz, G. Postimplantation expression patterns indicate a role for the mouse forkhead/HNF-3 alpha, beta and gamma genes in determination of the definitive endoderm, chordamesoderm and neuroectoderm. *Development (Cambridge, England)* **119**, 567-578 (1993).

140. Ruiz i Altaba, A., Prezioso, V., Darnell, J. & Jessell, T. Sequential expression of HNF-3 beta and HNF-3 alpha by embryonic organizing centers: the dorsal lip/node, notochord and floor plate. *Mechanisms of development* **44**, 91-108 (1993).

141. Weinstein, D. *et al.* The winged-helix transcription factor HNF-3 beta is required for notochord development in the mouse embryo. *Cell* **78**, 575-588 (1994).

142. Ang, S. & Rossant, J. HNF-3 beta is essential for node and notochord formation in mouse development. *Cell* **78**, 561-574 (1994).

143. Le Lay, J. & Kaestner, K. The Fox genes in the liver: from organogenesis to functional integration. *Physiological reviews* **90**, 1-22 (2010).

144. Bardot, E. *et al.* Foxa2 identifies a cardiac progenitor population with ventricular differentiation potential. *Nature communications* **8**, 14428 (2017).

145. Behr, R. *et al.* Mild nephrogenic diabetes insipidus caused by Foxa1 deficiency. *The Journal of biological chemistry* **279**, 41936-41941 (2004).

146. Kaestner, K., Katz, J., Liu, Y., Drucker, D. & Schütz, G. Inactivation of the winged helix transcription factor HNF3alpha affects glucose homeostasis and islet glucagon gene expression in vivo. *Genes & development* **13**, 495-504 (1999).

147. Sund, N. *et al.* Tissue-specific deletion of Foxa2 in pancreatic beta cells results in hyperinsulinemic hypoglycemia. *Genes & development* **15**, 1706-1715 (2001).

148. Lee, C., Sund, N., Behr, R., Herrera, P. & Kaestner, K. Foxa2 is required for the differentiation of pancreatic alpha-cells. *Developmental biology* **278**, 484-495 (2005).

149. Lee, C., Friedman, J., Fulmer, J. & Kaestner, K. The initiation of liver development is dependent on Foxa transcription factors. *Nature* **435**, 944-947 (2005).

150. Morris, S.A. Direct lineage reprogramming via pioneer factors; a detour through developmental gene regulatory networks. *Development (Cambridge, England)* **143**, 2696-2705 (2016).

151. Huang, P. *et al.* Direct reprogramming of human fibroblasts to functional and expandable hepatocytes. *Cell stem cell* **14**, 370-384 (2014).

152. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663-676 (2006).

153. Morris, S.A. *et al.* Dissecting engineered cell types and enhancing cell fate conversion via CellNet. *Cell* **158**, 889-902 (2014).

154. Mirny, L.A. Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 22534-22539 (2010).

155. Voss, T.C. *et al.* Dynamic Exchange at Regulatory Elements during Chromatin Remodeling Underlies Assisted Loading Mechanism. *Cell* **146**, 544-554 (2011).

156. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature genetics* **43**, 264-268 (2011).

157. Grøntved, L. *et al.* C/EBP maintains chromatin accessibility in liver and facilitates glucocorticoid receptor recruitment to steroid response elements. *The EMBO journal* **32**, 1568-1583 (2013).

158. Zhu, B. *et al.* Coactivator-Dependent Oscillation of Chromatin Accessibility Dictates Circadian Gene Amplitude via REV-ERB Loading. *Molecular Cell* **60**, 769-783 (2015).

159. Voss, T.C. & Hager, G.L. Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nature reviews. Genetics* **15**, 69-81 (2014).

160. 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).

161. Hota, S.K. & Bruneau, B.G. ATP-dependent chromatin remodeling during mammalian development. *Development (Cambridge, England)* **143**, 2882-2897 (2016).

162. Foxa2 and H2A.Z Mediate Nucleosome Depletion during Embryonic Stem Cell Differentiation. *Cell* **151**(2012).

163. Mizuguchi, G. *et al.* ATP-Driven Exchange of Histone H2AZ Variant Catalyzed by SWR1 Chromatin Remodeling Complex. *Science* **303**, 343-348 (2004).

164. Takaku, M. *et al.* GATA3-dependent cellular reprogramming requires activation-domain dependent recruitment of a chromatin remodeler. *Genome Biology* **17**, 36 (2016).

165. Gaspar-Maia, A. *et al.* Chd1 regulates open chromatin and pluripotency of embryonic stem cells. *Nature* **460**, 863-868 (2009).

166. Wang, L. *et al.* INO80 facilitates pluripotency gene activation in embryonic stem cell self-renewal, reprogramming, and blastocyst development. *Cell stem cell* **14**, 575-591 (2014).

167. Pardo, M. *et al.* An expanded Oct4 interaction network: implications for stem cell biology, development, and disease. *Cell stem cell* (2010).

168. Polo, J.M., Anderssen, E., Walsh, R.M. & Schwarz, B.A. A molecular roadmap of reprogramming somatic cells into iPS cells. *Cell* (2012).

169. Tuteja, G., Jensen, S.T., White, P. & Kaestner, K.H. Cis-regulatory modules in the mammalian liver: composition depends on strength of Foxa2 consensus site. *Nucleic acids research* **36**, 4149-4157 (2008).

170. Li, Z., Schug, J., Tuteja, G., White, P. & Kaestner, K.H. The nucleosome map of the mammalian liver. *Nature structural & molecular biology* **18**, 742-746 (2011).

171. Sung, M.-H.H., Guertin, M.J., Baek, S. & Hager, G.L. DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Molecular cell* **56**, 275-285 (2014).

172. Bailey, T.L. & Machanick, P. Inferring direct DNA binding from ChIP-seq. *Nucleic acids research* **40**(2012).

173. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-112 (2009).

174. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279-283 (2011).

175. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**, 576-589 (2010).

176. 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. *Nature methods* **10**, 1213-1218 (2013).

177. Zaret, K. Developmental competence of the gut endoderm: genetic potentiation by GATA and HNF3/fork head proteins. *Developmental biology* **209**, 1-10 (1999).

178. Yu, B. *et al.* Reprogramming fibroblasts into bipotential hepatic stem cells by defined factors. *Cell stem cell* **13**, 328-340 (2013).

179. Tsankov, A.M. *et al.* Transcription factor binding dynamics during human ES cell differentiation. *Nature* **518**, 344-349 (2015).

180. Jozwik, K.M., Chernukhin, I., Serandour, A.A., Nagarajan, S. & Carroll, J.S. FOXA1 Directs H3K4 Monomethylation at Enhancers via Recruitment of the Methyltransferase MLL3. *Cell reports* **17**, 2715-2723 (2016).

181. Brinkman, A.B. *et al.* Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome research* **22**, 1128-1138 (2012).

182. Rhee, H.S. & Pugh, B.F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408-1419 (2011).

183. Sekiya, T. & Zaret, K.S. Repression by Groucho/TLE/Grg proteins: genomic site recruitment generates compacted chromatin in vitro and impairs activator binding in vivo. *Molecular cell* **28**, 291-303 (2007).

184. Sakaue-Sawano, A. *et al.* Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell* **132**, 487-498 (2008).

185. Zhang, Y. *et al.* Nucleation of DNA repair factors by FOXA1 links DNA demethylation to transcriptional pioneering. *Nature genetics* **48**, 1003-1013 (2016).

186. Wapinski, O.L. *et al.* Hierarchical mechanisms for direct reprogramming of fibroblasts to neurons. *Cell* **155**, 621-635 (2013).

187. Fong, A.P. & Tapscott, S.J. Skeletal muscle programming and re-programming. *Current opinion in genetics & development* **23**, 568-573 (2013).

188. Chronis, C. *et al.* Cooperative Binding of Transcription Factors Orchestrates Reprogramming. *Cell* **168**, 442 (2017).

189. Reményi, A. *et al.* Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes & development* **17**, 2048-2059 (2003).

190. Dan, S., Kang, B., Duan, X. & Wang, Y.-J.J. A cell-free system toward deciphering the post-translational modification barcodes of Oct4 in different cellular contexts. *Biochemical and biophysical research communications* **456**, 714-720 (2015).

191. Jang, H. *et al.* O-GlcNAc regulates pluripotency and reprogramming by directly acting on core components of the pluripotency network. *Cell stem cell* **11**, 62-74 (2012).

192. Brumbaugh, J. *et al.* Phosphorylation regulates human OCT4. *Proceedings of the National Academy of Sciences* **109**, 7162-7168 (2012).

193. Myers, S.A. *et al.* SOX2 O-GlcNAcylation alters its protein-protein interactions and genomic occupancy to modulate gene expression in pluripotent cells. *eLife* **5**(2016).

194. Belaguli, N.S., Zhang, M., Brunicardi, C.F. & Berger, D.H. Forkhead Box Protein A2 (FOXA2) Protein Stability and Activity Are Regulated by Sumoylation. *PLoS ONE* **7**(2012).

195. van Gent, R. *et al.* SIRT1 Mediates FOXA2 Breakdown by Deacetylation in a Nutrient-Dependent Manner. *PLoS ONE* **9**(2014).

196. Landt, S.G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research* **22**, 1813-1831 (2012).

197. Morisaki, T., Müller, W.G., Golob, N., Mazza, D. & McNally, J.G. Single-molecule analysis of transcription factor binding at transcription sites in live cells. *Nature communications* **5**, 4456 (2014).

198. Choy, J.S. *et al.* DNA methylation increases nucleosome compaction and rigidity. *Journal of the American Chemical Society* **132**, 1782-1783 (2010).

199. You, J.S. *et al.* OCT4 establishes and maintains nucleosome-depleted regions that provide additional layers of epigenetic regulation of its target genes. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 14497-14502 (2011).

200. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science (New York, N.Y.)* **356**(2017).

201. Wang, T. *et al.* Subtelomeric hotspots of aberrant 5-hydroxymethylcytosine-mediated epigenetic modifications during reprogramming to pluripotency. *Nature cell biology* **15**, 700-711 (2013).

202. Dunican, D.S., Pennings, S. & Meehan, R.R. The CXXC-TET bridge — mind the methylation gap! *Cell Research* **23**, 973-974 (2013).

203. Kwon, A.T., Arenillas, D.J., Worsley Hunt, R. & Wasserman, W.W. oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets. *G3 (Bethesda, Md.)* **2**, 987-1002 (2012).

204. Jocelyn Charlton, T.D., Alexander Meissner Global delay in nascent strand DNA methylation. *Science*.

205. Alabert, C. *et al.* Nascent chromatin capture proteomics determines chromatin dynamics during DNA replication and identifies unknown fork components. *Nature Cell Biology* **16**, 281-293 (2014).

206. Petruk, S. *et al.* Delayed Accumulation of H3K27me3 on Nascent DNA Is Essential for Recruitment of Transcription Factors at Early Stages of Stem Cell Differentiation. *Molecular cell* **66**, 247-25700000 (2017).

207. Ramachandran, S. & Henikoff, S. Replicating nucleosomes. *Science Advances* **1**(2015).

208. Marzluff, W.F., Wagner, E.J. & Duronio, R.J. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nature Reviews Genetics* **9**, 843-854 (2008).

209. Ransom, M., Dennehey, B.K. & Tyler, J.K. Chaperoning Histones during DNA Replication and Repair. *Cell* **140**, 183-195 (2010).

210. Annunziato, A.T., Schindler, R.K., Thomas, C.A. & Seale, R.L. Dual nature of newly replicated chromatin. Evidence for nucleosomal and non-nucleosomal DNA at the site of native replication forks. *Journal of Biological Chemistry* **256**, 11880-11886 (1981).

211. Jackson, V. & Chalkley, R. A reevaluation of new histone deposition on replicating chromatin. *The Journal of biological chemistry* **256**, 5095-5103 (1981).

212. Ran, A.F. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* **8**, 2281-2308 (2013).

213. 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).

214. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology* **30**, 271-277 (2012).

215. Yang, Y.A. & Yu, J. Current perspectives on FOXA1 regulation of androgen receptor signaling and prostate cancer. *Genes & Diseases* **2**, 144-151 (2015).

216. Arenas, E. Foxa2: the rise and fall of dopamine neurons. *Cell stem cell* **2**, 110-112 (2008).

217. Stott, S.R.W. *et al.* Foxa1 and Foxa2 Are Required for the Maintenance of Dopaminergic Properties in Ventral Midbrain Neurons at Late Embryonic Stages. *The Journal of Neuroscience* **33**, 8022-8034 (2013).

218. Fernández‐Santiago, R. *et al.* Aberrant epigenome in iPSC‐derived dopaminergic neurons from Parkinson's disease patients. *EMBO Molecular Medicine* **7**, 1529-1546 (2015).

219. Zeng, X. & Couture, L.A. Pluripotent stem cells for Parkinson's disease: progress and challenges. *Stem Cell Research & Therapy* **4**, 25 (2013).

220. Zhang, P., Xia, N. & Pera, R.A. Directed dopaminergic neuron differentiation from human pluripotent stem cells. *Journal of visualized experiments : JoVE*, 51737 (2014).

221. Mikkelsen, T.S. *et al.* Comparative epigenomic analysis of murine and human adipogenesis. *Cell* **143**, 156-169 (2010).

222. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359 (2012).

223. Robinson, J.T. *et al.* Integrative genomics viewer. *Nature biotechnology* **29**, 24-26 (2011).

224. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**(2008).

225. Quinlan, A.R. Current Protocols in Bioinformatics. *wiley* (2014).