The epigenetic and transcriptional consequences of aberrant FoxC1 expression in Acute Myeloid Leukaemia

Liam Niall Gilding

A thesis submitted to the University of Birmingham for the degree of DOCTOR OF PHILOSOPHY

Institute of Biomedical Research
Birmingham Centre for Genome Biology
College of Medical and Dental Sciences
University of Birmingham

September 2017

Abstract

FOXC1 encodes a mesenchymal transcription factor that is not normally expressed in haematopoietic cells. However, recent studies from this and another laboratory demonstrated that *FOXC1* is inappropriately de-repressed in Acute Myeloid Leukaemia (AML). Through epigenomic profiling of primary AML samples, we showed that *FOXC1* was specifically upregulated in the aggressive FLT3-ITD subtype of AMLs, in parallel with the activation of FOX:E-box composite *cis*-regulatory elements. Furthermore, complementary studies from the Somervaille laboratory demonstrated that FoxC1 expression in AML was leukaemogenic by establishment of a monocyte differentiation block and enhancement of clonogenic potential.

Collectively, these data indicated that FoxC1 plays a critical role in leukaemogenesis, but the target genes and mechanisms by which this occurred were not known. To address this, we performed an integrative genome-wide analysis of FoxC1 binding, chromatin accessibility and gene expression in primary AML samples, *in vivo* models and cell lines. These studies revealed that FoxC1 acts to block normal myeloid differentiation by contributing to the widespread repression of differentiation-specific target genes. Critically, we identify *Meis2*, a proto-oncogene which collaborates with Hoxa9, as a putative direct target of FoxC1, providing compelling indications of a potential FoxC1-dependent oncogenic mechanism.

Acknowledgments

Firstly, I express my sincerest thanks to my supervisor, Professor Peter Cockerill. Over the last four years, he has been a tremendous source of knowledge, motivation and mentorship. I also thank my co-supervisor, Professor Constanze Bonifer.

I am extremely grateful for the generosity and encouragement provided by my collaborator, Professor Tim Somervaille at the CRUK Manchester Institute. Without access to his expertise and resources, this study would have been much the poorer.

I owe a special thanks to Dr Peter Keane. Peter was not only fantastic with bioinformatics support, but he is also a generous teacher. I am grateful he was always willing to share the secrets of his trade with a curious outsider, and hash out scientific ideas over a pint. Thanks also to the rest of the Cockerill-Bonifer group for training and making our lab a fun environment to work in.

I also want to thank my primary school science teacher, Andy Hollingworth ('Mr H'), for inspiring my earliest enthusiasm for practical science (and for demonstrating the utility of proper experimental controls!).

Finally, I thank my family and friends, particularly my best friend Matt and housemate Alice, for keeping me buoyant through the most challenging patches. But above all, I am most indebted to my parents and Catherine for their unwavering support and encouragement of all my endeavours, both in and out of the lab.

This work was supported by a Bloodwise Gordon Piller Studentship.

Table of Contents

List of Abbreviations

AML	Acute Myeloid Leukaemia
BLBC	Basal-like breast cancer
BM	Bone Marrow
C/EBP	CCAAT enhancer binding protein
CaI	Calcium ionophore A23187
CAPTURE	CRISPR affinity purification *in situ* of regulatory elements
cDNA	Complementary DNA
Ch1	Chromosome 1
Ch18	Chromosome 18
ChIP	Chromatin Immunoprecipitation
CLP	Common Lymphoid Progenitor
CMP	Common Myeloid Progenitor
CPSF	Cleavage and polyadenylation specificity factor
CRISPR	Clustered regularly interspaced short palindromic repeats
CT	Chromosome Territory
CTCF	CCCTC-binding factor
CUT&RUN	Cleavage Under Targets and Release Using Nuclease
DHS	DNase I-hypersensitive site
DNA	Deoxyribonucleic acid
EMSA	Electrophoretic Mobility Shift Assay
FC	Fold change
FHD	Forkhead DNA-binding domain
FISH	Fluorescent *in situ* hybridisation
FL	FLT3 ligand
FLT3	FMS-like tyrosine kinase 3
FLT3-ITD	FLT3 internal tandem duplication
FOX:E	FOX:E-box
FRAP	Fluorescence recovery after photobleaching
GFP	Green fluorescent protein
GMP	Granulocyte-Macrophage Progenitor
GR	Glucocorticoid receptor
HAT	Histone acetyltransferase
HDAC	Histone deacetylase
HSPCs	Haematopoietic stem and progenitor cells
ID	Inhibitory domain
IMDM	Iscove's Modified Dulbecco Medium
JM	Juxtamembrane Domain
LAD	Lamina-associated domain
LB	Lysogeny broth
LMPP	Lymphoid-Primed Multipotent Progenitor
LT-HSC	Long-term Haematopoietic Stem Cell
LZ	Leucine zipper
MACS	Magnetic-activated Cell Sorting
MBD	methyl-CG binding domain
MBP	methyl-CG binding protein
MEP	Megakaryocyte-Erythroid Progenitor
MMTV	Mouse Mammary Tumour Virus

NES	Nuclear export signal
NLS	Nuclear localisation signal
nt	Nucleotide
O.D.	Optical density
OSK	Oct4, Sox2, Klf4
PBMCs	Peripheral Blood Mononuclear Cells
PBSCs	Peripheral Blood Stem Cells
PCR	Polymerase Chain Reaction
PIC	Pre-initiation complex
PMA	Phorbol 12-myristate 13-acetate
Poly(A)	Polyadenylation
RNA	Ribonucleic acid
RNAPII	RNA Polymerase II
RPMI	Roswell Park Memorial Institute 1640 Medium
RT	Room temperature
SEC	Super Elongation Complex
sgRNA	single guide RNA
shRNA	Short hairpin RNA
siRNA	Small interfering RNA
SOC medium	Super Optimal broth with Catabolite repression
ST-HSC	Short-term Haematopoietic Stem Cell
Ta	Annealing Temperature
TAD	Topological-associated domain
TPM	Transcripts per million
TRD	Transcriptional repressor domain

Chapter 1: Introduction

1.1 Chromatin structure and function

The first recorded description of chromatin was made by the cytologist Walter Flemming in the late 1800s, who observed “stainable material” in the nuclei of cells treated with a basophilic dye. Indeed, the word ‘chromatin’ is derived from the Greek word for ‘colour’, ‘khroma’[1]. However, it was not until many years later that the diverse functions of chromatin became apparent.

Firstly, chromatin contains the entire nuclear genome, which is approximately 3 billion bases of DNA in humans. The genome exists in chromatin as part of a complex macromolecular structure containing RNA, histone and non-histone proteins. The tight association of the genome with chromatin proteins is what enables sufficient compaction to accommodate ~2 m of linear DNA sequence into a nucleus less than 10 μm in diameter[2].

Secondly, it emerged that chromatin is highly dynamic and intrinsically connected to essential processes throughout the cell cycle. For example, during metaphase, chromatin reaches maximal compaction as discrete chromosomes to ensure complete transmission of genetic material onto daughter cells during mitosis. Furthermore, during interphase, chromatin structure and organisation is intimately involved in the regulation of gene transcription, DNA replication and DNA repair[3].

1.1.1 The nucleosome

The lowest level of chromatin structure is the simplest repeating unit of DNA and histones, known as the nucleosome. Although it was known for some time that chromatin contained both DNA and proteins, it was not until X-ray diffraction studies by Kornberg in

the 1970s revealed that chromatin was at its simplest level comprised of DNA associated with a complex of histone proteins, termed the nucleosome[4]. Subsequent advances in nucleosome research culminated in high resolution crystal structures of the core nucleosome particle some 20 years later[5,6]. Based on these structures and other biochemical data, we now have a very detailed understanding of what constitutes the core nucleosome particle and how they are assembled.

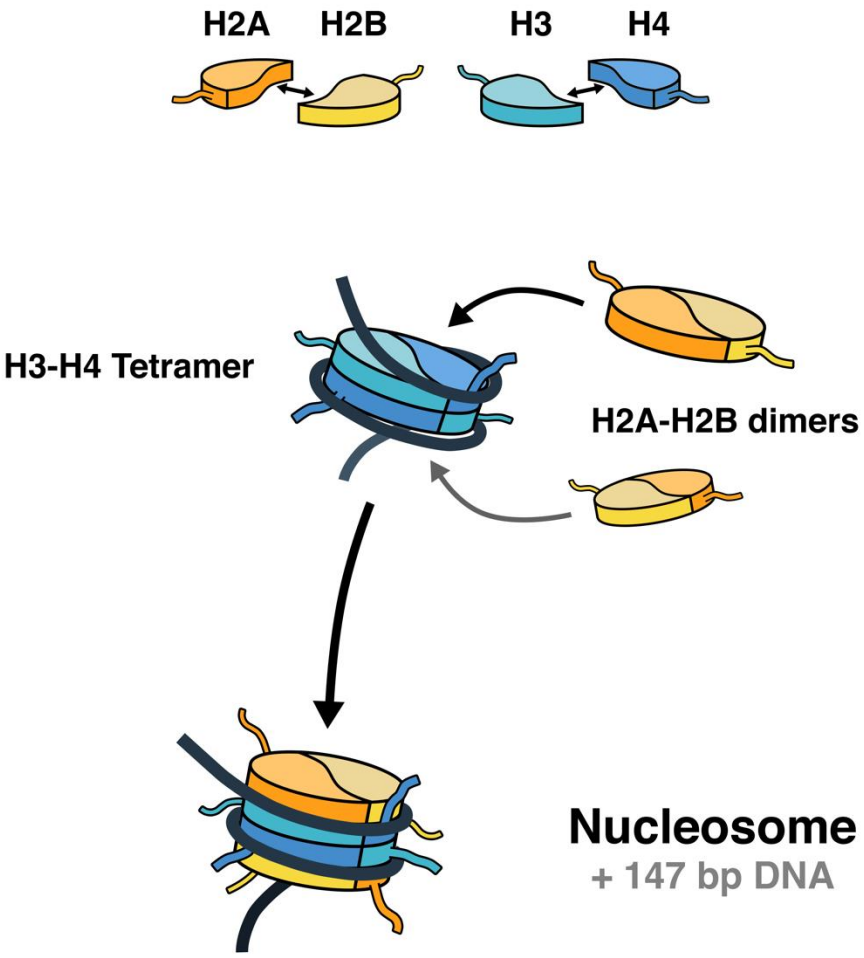


Figure 1:1 **Assembly of histone proteins into nucleosome particles.**

A schematic depiction of core histone assembly, where two H3-H4 dimers first associate to form a tetramer, followed by two H2A-H2B dimers above the H3-H4 tetramer[2].

Nucleosomes are comprised of an octameric complex of the core histones H2A/B, H3 and H4, which all feature a conserved structural helix-loop-helix domain termed the 'histone fold' (Figure 1:1). Two histone folds on discrete proteins form 'handshake motifs' which mediate head-to-tail heterodimerisation of histones H2A with H2B, and H3 with H4[6]. Two H3-H4 heterodimers form a tetramer which is bound above and below by H2A-H2B heterodimers to form the core histone octamer. Basic, arginine-rich grooves of individual histone monomers are presented on the outer surface of the octamer, forming a gyre-shaped channel which associates with the minor groove of DNA by electrostatic interations. In all, 147 bp of DNA is wound around the histone octamer ~1.65 times for a single core nucleosome particle[5].

1.1.2 Chromatin fibres

Multiple nucleosomes spontaneously assemble *in vitro* into long arrays separated by ~20-80 bp of intervening DNA to form chromatin fibres. These structures were originally visualised under salt-free conditions using electron microscopy and described as 10 nm 'beads on a string' arrays[7]. However, *in vitro* studies using physiological salt conditions demonstrated that nucleosome arrays condense further into 30 nm fibres, suggesting that the popular textbook images of 10 nm arrays very rarely exist *in vivo*, if at all[8]. Furthermore, experiments in living cells demonstrated that even the least condensed domains of chromatin existed as 30 nm fibres[9].

In native chromatin, two processes promote the compaction of nucleosomes into 30 nm fibres. Firstly, an additional linker protein, usually histone H1, associates with DNA entering nucleosomes to form larger particles incorporating an additional ~20 bp of DNA called chromatosomes[10]. Furthermore, recruitment of linker histones seems to increase the steric compliance of DNA, allowing tighter compaction of chromatin[11]. It was originally thought thought that these 30 nm fibres were organised into a simple solenoid, but it is

now generally assumed that chromatin assembles as an interdigitated zig-zag like structure[12,13].

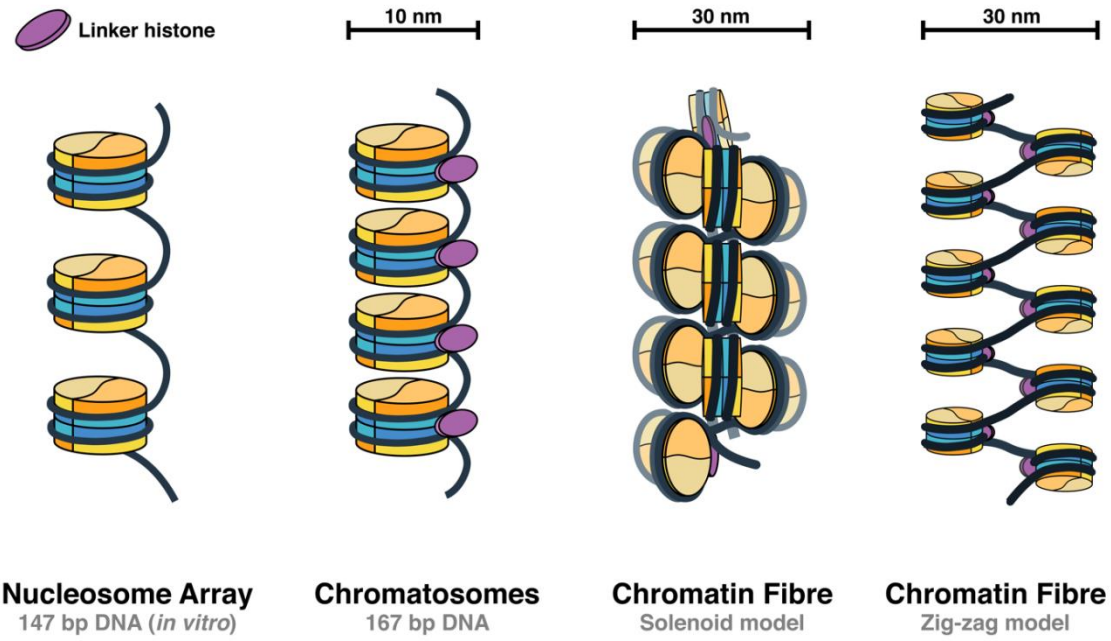


Figure 1:2 **Models of chromatin fibres.**

In vitro studies demonstrate nucleosomes spontaneously form regularly-spaced **nucleosome arrays**. However, *in vivo*, nucleosome arrays rarely exist, and are generally bound by linker histones to form more compact 10 nm **chromatosomes**. It is generally accepted that the next step in compaction is to form 30 nm chromatin fibres, however the precise structure acquired remains unclear. Both **solenoid** and **zig-zag** models of chromatin fibres have been proposed[12,13].

1.1.3 Chromatin domains

30 nm chromatin fibres containing several kilobases of DNA condense further into a hierarchy of discrete independently regulated chromatin domains. Genome-wide chromosome interaction studies revealed that the mammalian genome is organised into megabase-scale regions called 'topologically associated domains' (TADs). TADs are characterised by high chromatin interaction frequencies within the domain, but few interactions across different TADs[14,15]. The boundaries defining TADs are highly conserved, both across cell types in a given species and across evolution. However,

these TADs are differentially arranged in the nucleus according to cell-type specific gene activity[14]. Thus, changes in gene expression across different cell types arise not from changes in TAD boundaries, but rather the spatial repositioning of TADs between active and inactive chromosome compartments.

The boundaries of TADs are important sites for regulation of genome architecture, and are frequently bound by domain-organising proteins. These include CTCF, a DNA-binding transcription factor which in this context helps to bring linearly-separated segments of DNA into close proximity by looping, and functions at insulators to isolate gene activity within one TAD from another[14].

Finally, it is important to remember that the nuclear genome in eukaryotes does not exist as one continuous linear sequence of DNA in chromatin, but instead is composed of discrete chromosomes. These structures are most clearly visualised during mitosis, when chromatin reaches its most condensed state in metaphase, allowing individual chromosomes to be defined with microscopy[16]. Chromosomes and their domains both sit above the level of TADs in the hierarchy of genome organisation, and are also arranged according to gene activity.

1.1.4 Chromosomes occupy distinct territories within the nucleus

During interphase, chromosomes preferentially occupy discrete nuclear spaces called chromosome territories[17]. Chromatin domain localisation also correlates with transcription activity and gene density. Chromosomes are organised radially, with transcriptionally active chromatin domains preferentially localising to the centre of the nucleus, or into the inter-chromosomal spaces, while inactive chromatin domains tend to lie closer to the periphery (Figure 1:3). An illustrative example of this phenomenon in human cells is the gene-rich chromosome 19, which tends to be found closer to the

centre of the nucleus. By contrast, the gene-poor and inactive chromosome 18 is usually found nearer to the nuclear membrane[18].

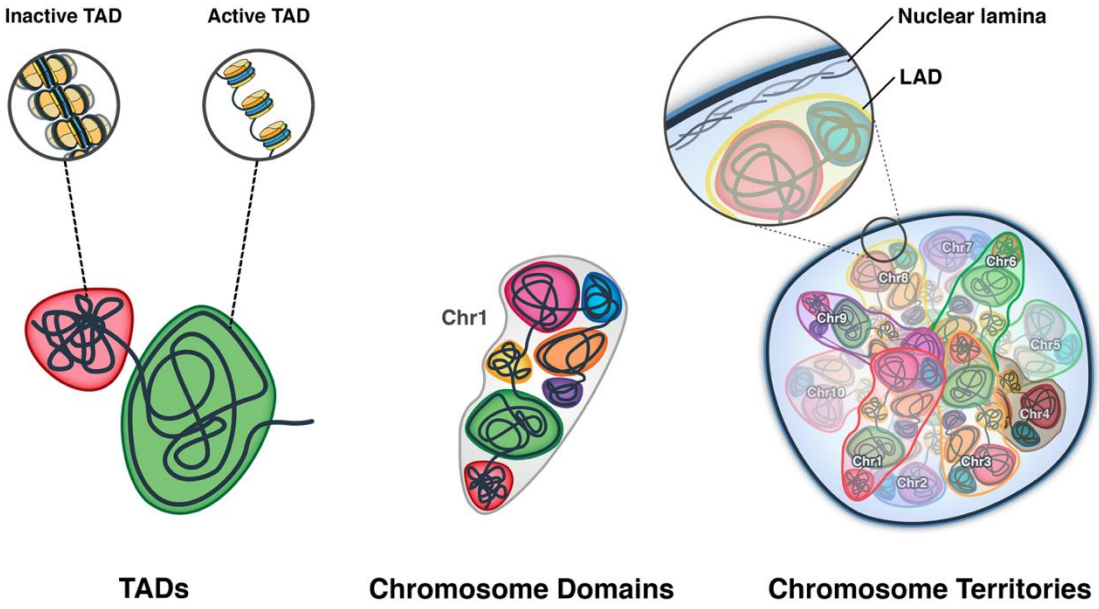


Figure 1:3 **Gemome organisation at the intra and inter-chromosomal level.**

TADs demarcate megabase level regions of chromosomes and are partitioned into chromosome domains in a tissue-specific manner, according to the activity of genes they contain. Furthermore, entire chromosomes are broadly organised within the nucleus according to their overall transcriptional activity[19]. The nuclear spaces preferentially occupied by particular chromosomes are referred to as chromosome territories. Chromosome domains that preferentially associate with the nuclear lamina are termed lamina-associated domains (LADs).

1.1.5 Chromatin is partitioned according to gene activity

Long before the identification of TADs in the late 2000s, it was already known that interphase chromatin exists in one of two general phases that broadly correlates with gene activity. The first such observations were made by Emil Heitz in 1928 when visualising moss nuclei under the microscope. He found that certain regions of chromosomes stained more intensely when treated with basophilic dyes and did not undergo decompaction after mitosis. These stably condensed regions were termed

'heterochromatin' and those that de-condensed following mitosis 'euchromatin'[20]. Since then, it is now known that these cytologically discrete states of chromatin are also functionally distinct: fluorescence *in situ* hybridisation (FISH), chromosome interaction assays and other approaches have all demonstrated that active chromosome domains are euchromatic in contrast to inactive domains, which exist as heterochromatin[14,21,22].

Several functional consequences arise from the partitioning of active and inactive chromatin domains. Firstly, the centre of the nucleus is enriched for active RNA polymerase and transcription factors, allowing robust expression of genes located in this space[19]. Secondly, the localisation of inactive heterochromatin domains to the periphery permits interaction of these domains with structural proteins of the nuclear lamina. These lamina-associated domains (LADs) contribute not only to the gross organisation of chromosomes in the nucleus by providing anchoring points, but are also associated with the stable repression of genes[23].

The spatial positioning of chromatin domains within the nucleus is not static, but is in fact a highly dynamic process intimately connected to shifts in gene expression patterns across cell types and developmental pathways. Domains containing inactive genes which become activated are repositioned from the nuclear periphery to the centre. Conversely, gene domains transitioning to a state of repression often move towards the outside of the nucleus. A prime example of these processes occurs at the *HoxB* cluster during mouse ES cell differentiation, where the induction of genes within the domain coincided with its repositioning towards the centre of the nucleus[24].

1.1.6 Chromatin is an intrinsic barrier to gene transcription

For gene transcription to occur, the underlying DNA sequence needs to be accessible to the basal transcriptional machinery to permit the recruitment and passage of RNA

polymerases. However, because of its highly condensed nature, chromatin is inherently repressive to transcription, as DNA sequences are not readily accessible to the transcriptional machinery. A critical step in gene activation is the localised de-compaction of chromatin at target genes and their associated *cis*-regulatory elements. Increased accessibility allows recognition of target sequences by transcription factors which, through various mechanisms discussed in 1.3.4, lead to gene activation[2].

1.1.7 Active chromatin is distinguished by increased nuclease sensitivity

A milestone in our understanding of active chromatin was made in the 1970s and 1980s by Weintraub and colleagues, who demonstrated that active genes resided in broad nuclease-sensitive domains[25,26]. In fact, active domains were found to be at least twice as sensitive to digestion by DNase I enzyme as inactive domains, with later studies suggesting nuclease sensitivity was perhaps an order of magnitude higher at specific loci[27]. In parallel studies, the nuclease sensitivity of particular chromatin domains was observed to be cell type-specific, and correlated with changes in local gene activity[28,29].

The heightened nuclease sensitivity of active loci is now accepted to arise from localised increases in chromatin accessibility which are necessary to allow recognition of target DNA sequences by transcription factors for gene activation[2]. Establishment of active chromatin also involves the deposition of covalent histone modifications and nucleosome remodelling, processes discussed in the next section. Collectively, these processes disrupt the interactions between histone H1 and linker DNA and the stable ordering of nucleosomes, causing localised de-compaction of chromatin. This exposes tracts of naked DNA for preferential binding by transcription factors *in vivo*, or digestion by nucleases *in vitro*. In addition to general de-compaction of active chromatin domains, there are also specific nucleosome-free genomic regions known as DNase I Hypersensitive Sites (DHSs) which exist at *cis*-regulatory elements (Figure 1:4). More

recently, genome-wide adaptations of chromatin analysis methods such as DNase-seq and ATAC-seq were developed and are used to map the entire complement of accessible chromatin features in specific cell types or conditions[30,31].

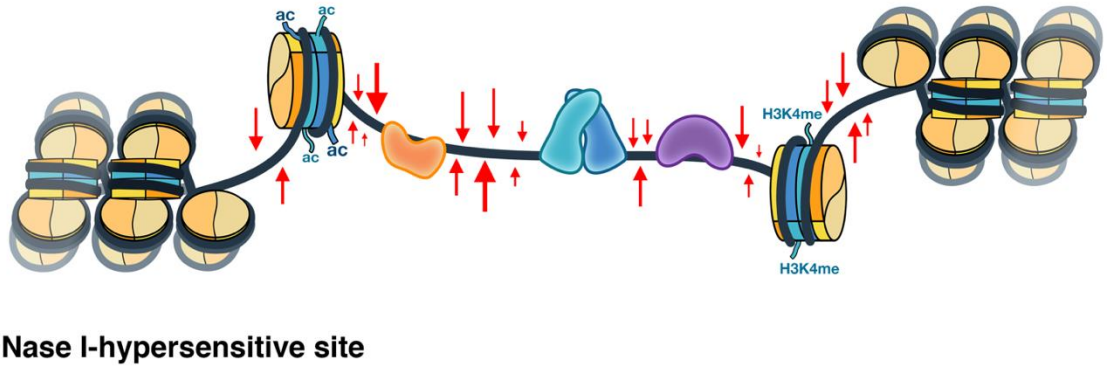


Figure 1:4 **Structure of DNase I-hypersensitive sites.**

Regions of accessible chromatin that generally correspond to active *cis*-regulatory elements feature exposed tracts of naked DNA which are preferentially sensitive to cleavage by DNase I and other nucleases (red arrows), in contrast to condensed chromatin to which DNase I enzyme cannot efficiently cleave.

Now that the general roles of chromatin in genome architecture and compaction have been introduced, the specific features of chromatin relating to gene activation shall be introduced.

1.2 Features of active and inactive chromatin

1.2.1 DNA methylation

The existence of methylated cytosine residues, or 5-methylcytosine (5mC), was first described in 1925, although a role for 5mC in gene activity only became clear some 50 years later[32,33]. DNA methyltransferases (DNMTs) catalyse the addition of methyl groups to cytosine, most commonly at CG dinucleotides to form 5mC. In mammals, symmetric methylation of both DNA strands is predominant, although asymmetric methylation also

occurs more rarely, particularly during germ-cell development[34,35]. DNA methylation occurs either as a replication-coupled process, preserving parental DNA methylation patterns on daughter strands, or in a replication-independent manner for *de novo* deposition[36]. DNA methylation dynamics begin to influence chromatin architecture and gene regulation almost immediately following fertilisation during embryogenesis[37].

Generally, 5mC deposition near a given gene is associated with its repression. Extensive 5mC deposition is also a feature of transcriptionally silenced pericentromeric and telomeric sequences, and during X-chromosome inactivation in female mammals[36]. The mechanisms of repression are not yet fully understood, but 5mC residues are generally thought to be targeted by methyl-CG binding proteins (MBPs) from the methyl-CG binding domain (MBD) or BTB/POZ families. MBPs directly impede the binding of gene-activating factors and recruit additional factors that establish stable gene repression[38,39].

In contrast to its deposition, removal of 5mC occurs by passive as well as active means. Passive demethylation most commonly arises from an absence of DNMT1 expression during DNA replication, leading to a two-fold dilution of methylated bases in cells with each division. The best example of passive methylation is during early embryogenesis, where low DNMT1 expression permits gradual removal of parental DNA methylation patterns, followed by *de novo* methylation catalysed by DNMT3A/B[35,40]. Active demethylation of 5mC is generally considered to be thermodynamically unfavourable, but it does occur as a by-product of single-stranded DNA repair processes[36].

Another, more recently described DNA modification is 5-hydroxymethylcytosine (5hmC), the product of 5mC oxidation by the ten-eleven translocation (Tet) family of oxygenases[41]. In contrast to 5mC, 5hmC enrichment at gene loci *positively* correlates

with their expression. Diverse roles for 5hmC both in normal development and disease are becoming apparent in this emerging field of intense study[42].

1.2.2 Covalent modification of histones and the histone code hypothesis

The earliest connection between histone modifications and gene regulation was made in the 1960s, when it was shown that histone acetylation increased gene transcription, and suggested that histone acetylation might exert locus-specific effects on gene regulation[43].

Since then, well over 100 different post-translational histone modifications - alone and in combination - have been catalogued and associated with a multitude of nuclear processes including gene regulation[44]. Furthermore, it has emerged that for many histone modifications, the genome-wide distributions are largely cell-type specific[45]. The histone code hypothesis, put forward by Allis and others in the early 2000s, postulates that key regulatory information relating to gene expression is embedded in locus-specific patterns of histone modificationss[46,47].

It is widely accepted that these reversible histone modifications play an important role in both chromatin organisation and gene expression, but the causal contribution they make to lineage-specific gene regulation is a topic of intense debate[2]. The view taken here is that many - but not all - histone modifications are a product of signal integration in chromatin by transcription factors and other proteins to increase the robustness of gene regulation, rather than intrinsic effectors of regulation themselves. Nevertheless, the mapping of histone modifications is a useful tool to identify the complement of *cis*-regulatory elements active in a given cell type. The most significant players in the 'histone code' as they feature in gene regulation (summarised in Table 1:1) are explored herein.

Table 1:1 **A summary of the principal histone modifications involved in the regulation of gene transcription.**

Modification	Associated with	Target histone residues
Acetylation	Activation	**H3** K9, K14, K18, K56 **H4** K5, K8, K12, K16 **H2B** K6, K7, K16, K17
Methylation	Activation	**H3** K4me2, K4me3 **H3** K36me3, K79me2
Methylation	Repression	**H3** K9me3, K27me3 **H4** K20me3

1.2.3 Histone acetylation correlates with gene activation

Following the connection between histone acetylation and gene activation being made by Allfrey in the 1960s, subsequent studies of the active chicken ß-globin locus demonstrated that histones in accessible regions were hyperacetylated[48]. In fact, of all the known histone modifications, acetylation exerts the most profound impact on chromatin accessibility. Acetylation of histone lysine residues disrupts the higher-order folding of the histone octamer, diminishing the electrostatic attractions between basic residues on histones and the phosphodiester backbone of DNA within nucleosomes, and also abrogating inter-nucleosomal interactions required for condensation into chromatin fibres. The resultant increase in chromatin accessibility enhances the recognition of underlying DNA sequences by transcription factors involved in gene regulation[2]. Of all the acetylation marks described, the two most relevant to gene activation are acetylation of histone H3 lysine 27 (H3K27ac) and H4 lysine 16 (H4K16ac).

Site-specific acetylation of histone lysine residues is catalysed by histone acetyltransferases (HATs), enzymes first shown to be directly involved in gene activation

in yeast[49]. Several mammalian orthologues have subsequently been identified[50]. Conversely, histone deacetylation is achieved by histone deacetylases (HDACs), causing localised re-compaction of chromatin and gene repression[51]. The best-characterised HATs in mammals are the p300/CBP and MOF proteins. Furthermore, lysine acetylation is also able to directly promote the recruitment of transcriptional activators that contain bromodomains that recognise various acetylated lysines on histones and other proteins[52].

p300 and CBP are closely related ubiquitously expressed co-activators with intrinsic HAT activity and a broad range of target histone lysine specificity[53], including H3K27. H3K27ac is usually enriched on nucleosomes flanking promoters and enhancers, and positively correlates with gene activity[54]. Mouse knockout studies have underlined the importance of these enzymes, where heterozygous deletion of either p300 or CBP led to severe viable phenotypes or embryonic lethality[55]. In addition to enhanced chromatin accessibility resulting from H3K27ac deposition, p300 and CBP are also thought to further promote gene activation by directly antagonising deposition of the repressive modification H3K27me3[56].

Finally, MOF catalyses deposition of H4K16ac, which has a particular significance in the establishment and maintenance of active chromatin. *In vitro* experiments have demonstrated that H4K16ac efficiently inhibits the condensation of nucleosome arrays into 30 nm fibres, and also disrupts the interactions between chromatin fibres that allow further compaction[57,58]. *In vivo*, H4K16ac is enriched throughout the bodies of active genes in human cells. In *Drosophila*, H4K16ac mediates gene-dosage compensation, a process leading to a global two-fold upregulation of gene expression[59,60].

1.2.4 The diverse effects of histone lysine methylation on gene activity

Histone lysine methylation was first reported in *Drosophila* to result from the activity of Su(var)3-9, and was associated with inactive heterochromatin[61]. Subsequent studies identified the human orthologue, SU(VAR)3-9, and the target residue as histone 3 lysine 9 (H3K9)[62,63]. H3K9 can be methylated up to three times to generate H3K9me1/2/3. Each methyl group provides a structural moiety that is recognised by repressive factors including heterochromatin protein 1 (HP1) which directly impairs nucleosome accessibility to inhibit transcription[64]. In fact, SU(VAR)3-9 belongs to a much broader class of enzymes, termed histone lysine methyltransferases (KMTs). A major subclass of KMTs comprising ~50 enzymes feature a common SET methyltransferase domain, with each enzyme targeting distinct lysine residues[65].

Of the multitude of other lysine methylation marks known, two further modifications of crucial relevance in gene regulation are H3 lysine 4 methylation (H3K4me) and H3K27 trimethylation (H3K27me3). In contrast to H3K9me, H3K4me is associated with transcriptional activation. Like H3K9me, KMT enzymes catalysing H3K4me1/2/3 also contain conserved SET domains[66]. Of all the H3K4 KMTs in mammals, the MLL proteins are the best characterised and have been shown in knockout studies to be essential genes. *Mll1* or *Mll2* heterozygous-null mice featured severe defects at birth, owing to comprehensive deregulation of homeotic genes involved in body patterning, while homozygous-null mice were embryonic lethal[67,68]. MLL proteins can catalyse the mono-, di- or tri-methylation of H3K4 to give H3K4me1/2/3, respectively[69].

H3K4me3 is an activating mark enriched at promoters, gene bodies and enhancer elements of active genes. At promoters, H3K4me3 is often maximally enriched most proximal to the transcription start site (TSS) compared to H3K4me1 and H3K4me2, typically at -300 bp and +100 bp[70]. Furthermore, at certain genes including tumour

suppressors, H3K4me3 enrichment exists as broad domains extending well into the gene body and is associated with increased transcriptional elongation. In the same study, broad domains of H3K4me3 were also reported at highly active enhancer elements[71]. Conversely, shortening of these broad H3K4me3 domains was indicative of diminishing transcription. H3K4me3 promotes gene activation via recruitment of the SWI/SNF chromatin remodelling complex. SWI/SNF promote lysine acetylation and nucleosome eviction, increasing accessibility to positively regulate transcription[72,73]. Although the above evidence would suggest that H3K4me3 is simply an activating mark, one study found that 59% of inactive gene promoters were also enriched for H3K4me3[70]. Partly accounting for conflicting reports regarding H3K4me3, studies by the Lander laboratory identified a distinctive pattern of chromatin at promoters of silent or lowly expressed genes featuring co-association of H3K4me3 with the repressive H3 lysine 27 trimethylation (H3K27me3) mark. They termed these promoters 'bivalent'[74].

H3K4me1 and H3K4me2 enrichment occurs at active enhancer elements in many tissue-specific contexts, and typically co-occurs with the activating H3K27ac mark[75]. Furthermore, H3K4me1 is a target for further activating proteins. For example, in the context of estrogen receptor alpha (ERα) signalling, ERα promotes the recruitment of the TIP60 histone remodelling enzyme to H3K4me1. TIP60 deposits an additional activating mark, H2A lysine 5 acetylation (H2AK5ac), at ERα target loci, further stimulating gene activation[76]. H3K4me1 is also elevated in the promoter regions of active genes together with H3K4me2/3[70].

In contrast to H3K4me1, the relationship between H3K4me2 and gene activation is more nuanced. Like H3K4me1, H3K4me2 is often enriched at active gene promoters, but is maximally enriched more proximal to the TSS than H3K4me1[70]. However, H3K4me2 enrichment also occurs at silent gene promoters in the absence of H3K4me3

(H3K4me2+/me3-). In one study of haematopoietic cells, silent genes in multipotent progenitors featuring H3K4me2+/me3- were rapidly activated when these cells differentiated, with an accompanying deposition of H3K4me3[77]. These and similar observations in other systems have led to the definition of H3K4me2+/me3- as a marker of transcriptional 'poising'. Further underlying its special role in gene regulation, H3K4me2 at tissue-specific enhancer regions is required for the recruitment of the pioneer factor FoxA1[78], which promotes target gene activation through mechanisms discussed in more detail in 1.3.6.

1.2.5 Bivalent chromatin marks genes for later activation or repression

Bivalent chromatin domains were identified from ChIP-Chip studies of H3K4me3 and H3K27me3, which found the individual patterns overlapped at a substantial number of silent promoters of genes in mouse embryonic stem (ES) cells. While H3K4me3 was mostly found proximal to the silent TSS, H3K27me3 marked broader domains across the TSS. Furthermore, re-ChIP experiments confirmed that nucleosomes were often co-occupied by these modifications[74].

The significance of bivalent promoters was realised when it was observed that they often regulated critical developmental transcription factor genes. An elegant example of this is the neural-specific regulator *Olig1*, which has a bivalent promoter in mouse ES cells. Upon differentiation into neural progenitor cells (NPCs), *Olig1* was strongly upregulated and the repressive H3K27me3 mark was lost, whilst the activating H3K4me3 was maintained[74]. By contrast, the non-neural *Gata6* locus is bivalently marked and lowly expressed in ES cells, and differentiation into NPCs silenced *Gata6* and led to loss of H3K4me3. However, *Gata6* expression is upregulated in macrophage development, and in these cells the bivalent promoter switched instead to form poised to fully-activated (H3K4me3+ve/H3K27me-ve) state[79].

H3K27me3 deposition is catalysed by multi-subunit Polycomb 2 (PRC2) complexes, which are thought to be recruited to target promoters by several features including lncRNA transcription and high levels of DNA methylation[80,81]. PRC2 contains a catalytic subunit, EZH2, with a conserved SET domain responsible for H3K27 KMT activity. Another PRC2 co-factor, Jarid2, aids in targeting the complex to chromatin and induces an allosteric change in EZH2 to promote H3K27me3 depostion, leading to localised nucleosome condensation[80]. *In vitro* studies have shown a single PRC2 complex compacts about three nucleosomes via interaction with the histone core[82]. This is a silencing mechanism distinct from H3K9me marking discussed in 1.2.4, which promotes recruitment of repressive HP1 complexes[63].

Finally, PRC2 occupancy competes with SWI/SNF binding at bivalent chromatin[83,84]. This is just one example of a tightly-regulated functional antagonism between activating and repressive factors at bivalent chromatin. Subtle shifts in the balance between these factors, for example in response to developmental stimuli, can lead to targeted tissue-specific gene expression. Critically, perturbation of this antagonistic relationship is a recurring feature of oncogenic transformation[85].

1.2.6 Histone variants

In addition to post-translational modifications of canonical histones, several histone variants with alternative amino acid sequences also exist[86]. Histone variants often play even more dynamic roles than their canonical counterparts. For example, although canonical histones are almost exclusively synthesised during S-phase of the cell cycle and are incorporated into nucleosomes in a replication-dependent manner, histone variants are expressed and incorporated throughout the cell cycle. Histone variants are recruited to target nucleosomes by a complement of histone chaperone proteins which mediate histone exchange[86].

The divergent amino acid sequences of histone variants affects their tertiary structure, impacting both nucleosome stability and the strength of DNA:histone interactions, leading to localised changes in chromatin compaction. Of particular relevance to gene activation are the variants H2A.Z and H3.3, which are enriched in accessible chromatin[87]. For example, nucleosomes containing H3.3 are more sensitive to disruption than those with canonical H3 alone, whereas nucleosomes containing H2A.Z do not efficiently recruit linker histones for further chromatin compaction[88]. By contrast, the macroH2A variant stabilises nucleosomal structure and has a repressive effect on transcription[89]. MacroH2A is often localised at silenced genes and on portions of the inactivated X chromosome in mammals[90].

1.2.7 ATP-dependent chromatin remodellers

In addition to specific enrichment of histone modifications and histone variants, active regions of the genome also feature highly dynamic nucleosome positioning. Re-positioning of nucleosomes is an ATP-dependent process driven by large multi-subunit complexes of chromatin remodelling factors. The best characterised chromatin remodellers belong to the SWI/SNF, IWSI and SWR1 families, which are grouped according to the distinct ways in which they alter nucleosome conformation[91]. For example, SWI/SNF complexes directly evict nucleosomes, while IWSI complexes slide nucleosomes along the DNA spooled around it. In gene activation, the consequence of both SWI/SNF and IWSI family members is the creation of nucleosome-free regions that expose target DNA sequences for binding by transcription factors[2,91]. SWI/SNF recruitment to target genes can be a transcription factor-dependent process, or it can be via intrinsic bromodomain-mediated recognition of acetylated histones[92].

ISWI can also repress gene activation by sliding nucleosomes over TSS of genes, occluding their core promoter elements for recognition by *trans*-acting factors. SWR1

families catalyse the substitution of canonical H2A-H2B histone dimers for those containing the H2A.Z variant. Nucleosomes containing H2A.Z are often enriched at active gene promoters and the intron-exon boundaries of actively transcribed genes[93].

1.3 Regulation of Gene Expression

A unifying theme of all the processes discussed so far is the transition of chromatin between a condensed, inactive state and a more relaxed, accessible structure which favours gene transcription. Now that these properties have been introduced, the mechanisms by which genes are precisely regulated shall be explored.

1.3.1 Mechanisms of gene transcription by RNA polymerase II

Transcription of protein-coding genes in eukaryotes is co-ordinated chiefly by a large RNA polymerase II (RNAPII) complex. The RBP1 subunit of RNAPII features a C-terminal domain (CTD) containing 52 tandem repeats (in humans) of the consensus sequence Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7, which are extensively regulated by phosphorylation throughout the course of transcription. RNAPII works in concert with several other proteins including general transcription factors (GTFs), activators and co-activators[94].

During the first phase of transcription, termed 'initiation', RNAPII is recruited to the DNA. This begins with the co-operative assembly of GTFs on to the nucleosome-free core promoter sequence, usually starting with TFIID. This complement of factors is known as the pre-initiation complex (PIC) which permits recruitment of RNAPII and other cofactors immediately upstream of the TSS[95]. In the presence of ATP, the core DNA sequence bound by the PIC is melted to produce a 'transcription bubble' and the open promoter complex[94] (Figure 1:5). Formation of the open promoter complex brings the template strand of DNA into close proximity with the RNAPII active site, which begins nascent RNA synthesis. In this context, the CTD of RNAPII is largely dephosphorylated[96].

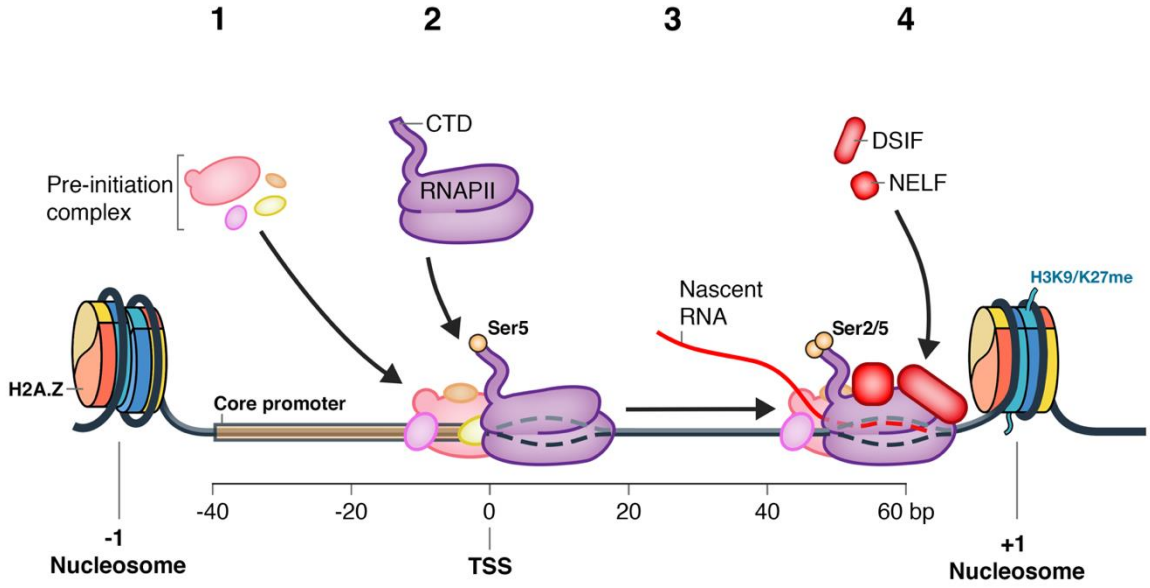


Figure 1:5 **Initiation of transcription at the promoter proximal region**

1: Assembly of the pre-initiation complex at the TSS; **2:** Recruitment of RNAPII to the core promoter; **3:** Initial burst of transcription; **4:** The presence of the +1 nucleosome and binding of the pausing factors DSIF and NELF to RNAPII lead to pausing of polymerase complexes prior to the elongation phase of transcription.

Adapted from Chen *et al.* (2015).

Following recruitment to the core promoter during initiation, RNAPII enters the elongation phase of transcription. After an initial burst of 20 to 60 nucleotides of RNA synthesis over the TSS, RNAPII frequently pauses. Precise mechanisms for pausing have yet to be determined, but it is thought that the +1 nucleosome immediately downstream of the TSS presents a physical barrier hindering RNAPII elongation[97]. Furthermore, the presence of pausing factors including PIC subunits and others including negative elongation factor (NELF) seem to tether RNAPII, and are displaced during 'escape' of RNAPII from the promoter-proximal region[98]. At highly-inducible genes, it was originally thought that multiple paused RNAPII complexes accumulate at the promoter-proximal region[99]. However, a recent high resolution footprinting study of RNAPII binding dynamics in

single cells proposed an alternative pausing mechanism featuring rapid RNAPII turnover at the promoter-proximal regions of active or inducible genes[100].

Escape of RNAPII from the promoter-proximal region is dependent upon phosphorylation of the paused RNAPII CTD, a process mediated mainly by BRD4 and the P-TEFb complex. In actively cycling cells with abundant gene transcription, up to 90% of P-TEFb is inactivated by its association with HEXIM, 7SK small nuclear RNA and other components. However, recruitment of BRD4 to promoters during gene activation outcompetes binding of these inhibitory factors to P-TEFb, activating the CDK9 kinase domain of P-TEFb[101]. Activated P-TEFb phosphorylates Ser5 of the RNAPII CTD, evicts the pausing factor NELF, and phosphorylates a target residue of another pausing factor, DSIF, which becomes permissive to elongation. The elongation phase begins with RNAPII Ser2 phosphorylation and the eviction of PAF1 from RNAPII, presenting binding surfaces on the polymerase complex that promote assembly of the super elongation complex (SEC)[102]. Assembly of the SEC aids in the transition to productive elongation by recruiting chaperones and the FACT complex, causing eviction of H2A-H2B dimers from downstream nucleosomes. This allows RNAPII to efficiently transcribe from DNA without displacement of remaining core histones in the nucleosome. Complete histone octamers rapidly reform following RNAPII procession, owing to a localised pool of free H2A-H2B dimers and the activity of histone chaperones[103,104]. Productive elongation by RNAPII is initially quite inefficient, with a transcription rate of ~0.5 kb/min. This rises to 2-5 kb/min after ~15 kb, as RNAPII accumulates Tyr1 and Ser2 phosphorylation and pausing factors across the gene body are removed[105].

Finally, as RNAPII approaches the 3' end of a gene and the polyadenylation (Poly(A)) site, Tyr1 residues in the RNAPII CTD are rapidly dephosphorylated, prompting recruitment of cleavage and polyadenylation specificity factor (CPSF) complex. CPSF

simultaneously interacts with RNAPII and the Poly(A) site to stall elongation and induce pausing, leading to the generation of R-loops forming between the nascent transcript and the genome. R-loop structures are resolved by SETX which permits access of the exoribonuclease XRN2 to degrade the 3' end of the nascent transcript, which is thought to eject RNAPII from the 3' end of the gene, terminating transcription[106,107]. The CPSF complex cleaves the pre-mRNA, permitting its release for subsequent processing and maturation.

1.3.2 The transcriptional machinery persists at active genes

At sporadically or lowly expressed genes, transcriptional termination is followed by the gradual reacquisition of an ordered nucleosome array across the gene, re-compaction of chromatin and release of the transcriptional machinery[108]. In contrast, active genes feature conformational differences that prime continued transcription. Firstly, 'scaffold PICs' persist at the promoter-proximal region even after an RNAPII complex escapes to begin productive elongation. Scaffold PICs promote high levels of gene expression by facilitating multiple successive rounds of RNAPII recruitment and transcription[109,110].

Finally, multiple actively-transcribed genes co-localise, appearing as visible foci using microscopy. These regions, distinguished by a concentrated pool of active RNAPII and associated factors, were termed 'transcription factories'[111] (Figure 1:6). Furthermore, genes localised to a given transcription factory are often regulated by similar combinations of sequence-specific transcription factors. In contrast to other models of transcription, it is thought that RNAPII complexes in transcription factories are spatially fixed, with the template DNA being fed through the transcribing RNAPII by other mechanisms[112]. Finally, genes can be moved in and out of transcription factories, providing another layer of regulation[113].

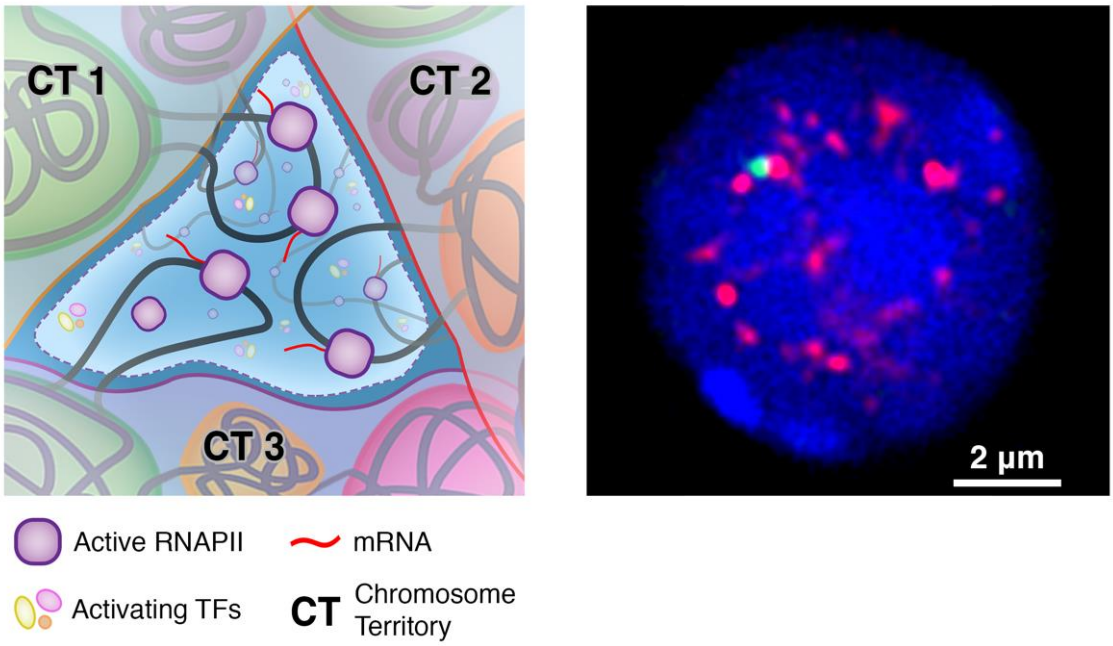


Figure 1:6 **Transcription factories**. (Left) Highly active regions of chromosomes co-localise within nuclear spaces enriched for active RNAPII complexes and activating co-factors, called transcription factories. (Right) These regions can be visualised under microscopy as distinct foci in the nucleus, using a combination of DNA FISH (green) and immunofluorescence targeting active RNAPII (red). Image on right from Bender *et al.* (2012).

1.3.3 Tissue-specific gene regulation is driven by transcription factors which shape the epigenetic landscapes of different cell types

An inherent characteristic of multicellular eukaryotes is the vast range of different cell types that make up the body, yet almost all of these cells share a common genome. To generate this functional diversity, genes need to be transcribed at the right level, in the right place and at the right time. Waddington was the first to define epigenetics as the study of processes "by which the genes of the genotype bring about phenotypic effects", underlining longstanding observations that the specialised functions of different cell types in complex organisms could not be explained by Mendelian genetics alone[114]. He later proposed that the phenotypic diversity of differentiated cells arise not from alterations in genetic inheritance, but instead from changes in the "epigenetic landscape" during organismal development[115] (Figure 1:7).

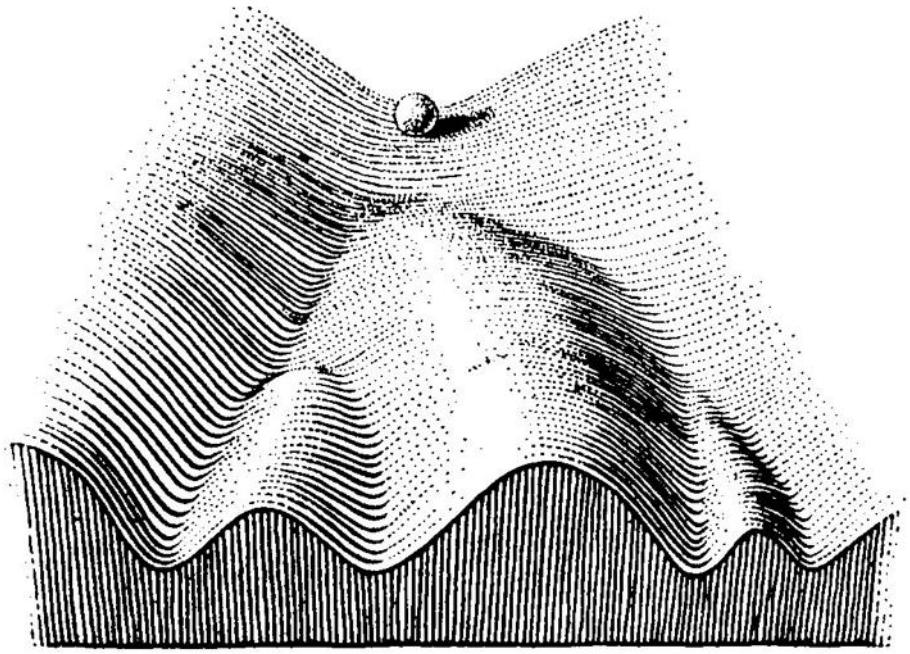


Figure 1:7 **Waddington's epigenetic landscape.** This classical model depicts a cell as a ball at the top of a hill, with the potential to follow one of multiple paths according to subsequent fate restriction events (depicted as branches/valleys in the hill). However, once the cell has passed a certain point, its fate is determined, and it cannot travel back up the hierarchy to a more 'potent' state[114,115].

Shortly after Waddington's epigenetic landscape model was proposed, Jacob and Monod found empirical evidence of the relationship between environment and gene expression as being central to changes in phenotype. In pioneering studies, they observed that genes could be induced or repressed via "functionally specialised genetic determinants" in response to "specific metabolites"[116]. Today, we know of these "genetic determinants" by another name: lineage-specific transcription factors (TFs), distinct from the GTFs which are involved in the transcription of all genes.

TFs work in concert at *cis*-regulatory elements of target genes to promote, maintain or repress the active chromatin structure required for gene transcription. Individual TFs form nodes in highly interconnected regulatory networks, with 'master regulator' TFs exerting the most influence in these expression control circuits. The first master regulators to be identified were the 'bithorax' genes, which emerged in the 1960s and 1970s from mutant

studies in *Drosophila* as essential for correct morphogenesis and segmentation in larval development. Disruption of any of these genes leads to a range of abnormal body plan phenotypes[117]. Bithorax genes were found to be part of a much broader class of 'homeotic' TF genes, and orthologues were subsequently identified in vertebrates. As in *Drosophila*, disruption to these genes, particularly those in the *HOX* gene family, result in severe developmental defects in both mice and humans[118,119].

The potential of master regulator TFs to fully specify developmental pathways is also elegantly illustrated by cellular reprogramming studies, where cell lineages can be switched from one type to another by the forced expression of just one or a select few TF genes. For example, over-expression of the myogenic master regulator *MyoD* is sufficient to reprogram adult fibroblast cells into muscle cells[120]. Furthermore, Shinya Yamanaka demonstrated that the expression of four TF genes alone could convert fully differentiated somatic cells into induced pluripotent stem cells (iPSCs) with reasonable efficiency[121].

1.3.4 Generalised roles and properties of transcription factors

The primary roles of all TFs, including the GTFs introduced in 1.3.1, are to control the rate of gene transcription. Distinct from GTFs, TFs integrate changes in cellular signalling and metabolism to elicit targeted regulation of specific genes. For this reason, TF proteins feature regulatory sites that are post-translationally modified by signalling proteins, or domains which bind metabolic co-factors, nuclear hormones or other TFs. TFs often belong to large families related by a common DNA-binding domain, for example the Forkhead (Fox) family[122]. Consequently, different members of such TF families may target similar DNA sequences, but have diverse effects on transcription due to differences in functional domains or interacting partners.

The broadest distinction between TFs was made based on their observed effect of increasing (activator TFs) or decreasing (repressor TFs) transcription. However, all TFs and co-factors alter transcription by the following principle mechanisms:

1. Interaction with target gene promoters to affect the assembly of the PIC and RNAPII recruitment at the promoter-proximal region[123].

2. Modulation of the activity of the general transcriptional machinery by catalysing post-translational modifications, conformational changes or recruitment/eviction of interacting partners[124,125].

3. Activation or repression of the activity of chromatin remodellers to open, maintain or close chromatin at target loci[126].

4. Promoting the physical interaction of distal *cis*-regulatory elements with the promoters of target genes to to stimulate activation, or to partition other genes into inactive chromatin domains[127,128].

The majority of TFs bind directly with target DNA sequences via specialised DNA-binding domains. The primary amino acid sequence and tertiary structure of these domains largely determine the specific DNA sequences, or binding motifs, which are recognised by particular TFs. Furthermore, many TFs feature interaction domains which allow binding to DNA motifs as multimeric complexes. One such example is the 'basic region leucine zipper' (bZIP) family of TFs. bZIP monomers dimerise with their leucine zipper domains prior to interacting with DNA via positively charged arginine and lysine residues in the basic region. In many cases, the monomers comprising a functional TF complex can be the same protein (homotypic) or different (heterotypic), creating functional diversity that either increases specificity of DNA binding or provide redundancy for robust regulation of critical genes[129].

Co-activators and co-repressors are distinct from TFs in that they do not directly bind DNA, but instead modulate the activity of DNA-binding TFs by other means. One class can co-operatively promote the PIC assembly at promoters, as is the case of the TBP associating factors (TAFs). TAFs recruit subunits of the basal transcriptional machinery to sequence-specific TFs bound at promoters via several interaction domains[130]. Another class of co-factors affect transcription through their intrinsic histone modifying activity. For example, the HAT enzymes CBP and p300, previously discussed in 1.2.7, are often recruited by TFs bound at target loci whereupon they acetylate adjacent histones to stimulate gene activity[131].

1.3.5 Transcription factors co-operate to disrupt nucleosomes at target loci

Despite their evident roles in gene regulation, the overwhelming majority of TFs cannot efficiently recognise binding sites on DNA if they reside on nucleosomes in chromatin. However, there is considerable evidence to show that TFs co-operate to bind target sequences which are otherwise occluded by their association with nucleosomes[132,133]. This mechanism was elegantly demonstrated in pioneering *in vitro* studies by the Workman laboratory. They found that three factors - GAL4, USF and NF-κB – individually bind inefficiently to intra-nucleosomal target sites located 46 bp from the nucleosome boundary. However, if these target sites were located within 20 bp of the nucleosome boundary, individual binding was significantly more efficient for all of these TFs. The +20 bp site is more efficiently bound by individual TFs because the association between DNA and histones is much weaker at nucleosome boundaries than at intranucleosomal sequences. Furthermore, the binding of any of these factors at the +20 bp site greatly increased the ability of TFs to consequently bind to the +46 bp nucleosomal site, by as much as 2 orders of magnitude. This was significant because USF and NF-κB could both

subsequently bind to intranucleosomal sequences, demonstrating an intrinsic cooperativity of binding.

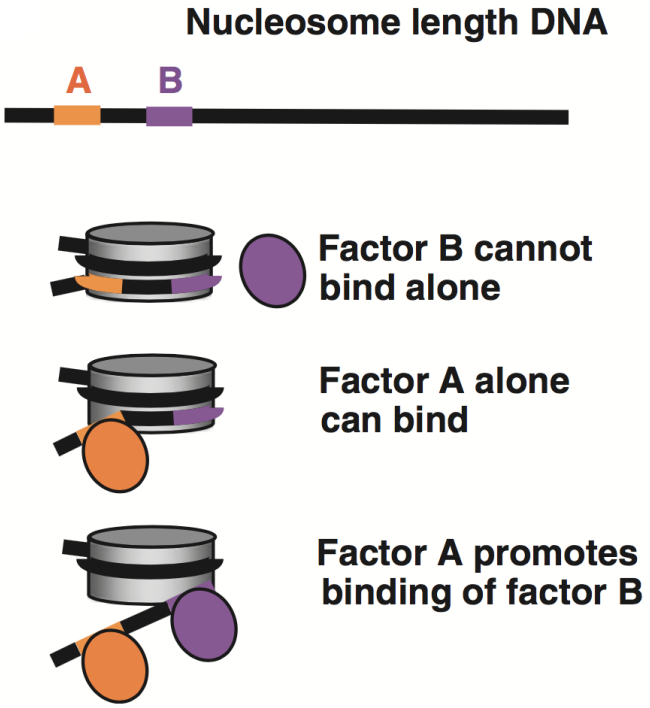


Figure 1:8 **Transcription factors co-operate to bind to nucleosomal sites.** A generalised model from Cockerill (2011), in which the binding of factor A to a partially exposed binding site at the nucleosome boundary sufficiently dissociates adjacent nucleosomal DNA for a second factor to subsequently bind.

The early *in vitro* studies which established the above model of TF binding have been supported more recently by genome-wide and live cell approaches, which confirm co-operative binding as a broader mechanism of TF recruitment to *cis*-regulatory elements[134,135]. A generalised model of co-operative binding is depicted in Figure 1:8, where the initial binding of factor A to a site at the nucleosome boundary disrupts weak DNA:histone interactions, partially 'unzipping' DNA from the nucleosome and allowing subsequent recognition of sites by factors B and then C. In this manner, individual TFs with little or no nucleosome binding properties can co-operate to progressively

destabilise nucleosomes and activate target *cis*-regulatory elements[2]. Critically, the vast majority of these TFs can still not recognise target binding motifs as they exist in condensed chromatin. In this context, DNA sequences that are not wrapped around the core nucleosome are either occluded by association with histone H1/H5 or neighbouring nucleosomes in chromatin fibres, preventing the initial binding event for co-operative binding to occur[136]. Prior events that increase chromatin accessibility – particularly the eviction of histone H1, which disrupts higher order nucleosome folding – are still essential pre-requisites for the overwhelming majority of TFs to recognise binding sequences and for co-operative binding to follow[137].

1.3.6 The specialised properties of pioneer factors in gene regulation

Pioneer factors are distinguished from all other TFs by their unique ability to efficiently bind target sites in condensed chromatin fibres *de novo,* without prior de-compaction by other processes. Pioneer factors such as the Forkhead domain (FOX) TFs gain entry into chromatin fibres by outcompeting binding of linker histones H1 or H5 via specialised protein domains, increasing local chromatin accessibility for other factors to recognise and bind their binding sequences. As a result, pioneer factors promote gene activation by conferring 'transcriptional competence' at target loci.

Establishment of transcriptional competence by pioneer factors can be either an active or a passive process (Figure 1:9). In the active role, pioneer factors target sites on compacted chromatin and rapidly increase local accessibility, exposing other sequences for recognition by conventional TFs[137]. Conversely, passive regulation involves stable pioneer factor occupancy at accessible chromatin to prime *cis*-elements for rapid induction by promoting the subsequent assembly of other factors. Excellent examples of passive pioneering are observed during androgen and estrogen receptor activation, which are often recruited to *cis*-elements 'bookmarked' by the pioneer factor FoxA1[138,139].

Finally, pioneer factors can persist at target sites throughout the cell cycle, ensuring rapid re-induction of crucial genes following mitosis[140].

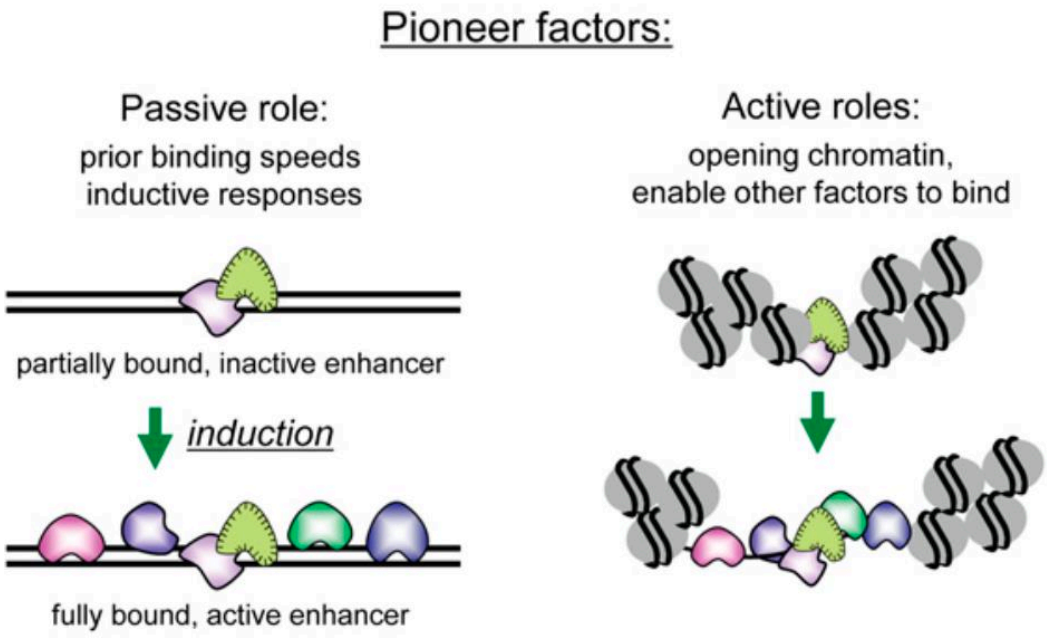


Figure 1:9 **Pioneer factors can behave in a passive or active manner to prime gene activation.** In the passive context (left), the pioneer factor remains stably bound at inactive enhancer regions, priming these elements for subsequent activation in the presence of additional non-pioneer TFs, perhaps by maintaining an accessible chromatin structure. Conversely, in the active context (right), pioneer factors gain direct entry into compacted chromatin fibres, allowing other TFs to recognise target binding sites in the accessible region created.

Our current knowledge of how pioneer factors work is largely due to landmark studies from the Zaret laboratory on the prototypical pioneer factor FoxA1. They first identified FoxA1 as a liver-specific TF that could efficiently bind a target site in the Albumin gene enhancer, even though this sequence was located on the nucleosomal surface and not be easily recognised by conventional TFs[141]. Subsequently, they demonstrated that FoxA1 could bind target sites on nucleosomes more stably than on free DNA, in stark contrast to most other TFs[142]. Structural characterisation of FoxA1 identified the 'winged-helix' DNA-binding domain responsible for this specialised activity[143]. The winged-helix

domain closely resembles histone H1/H5, allowing FoxA1 to efficiently recognise target sites in compacted chromatin by directly outcompeting linker histone binding to DNA[144]. Point mutations to the winged-helix domain demonstrated that FoxA1 actually recognises sites in chromatin via both nonspecific nucleosome binding and sequence-specific DNA binding activities[140]. The unusual ability of FoxA1 to bind nucleosomes in a nonspecific manner greatly slows its nuclear mobility compared to conventional TFs, allowing it to 'scan' chromatin for target sequences[145] (Figure 1:10).

FoxA1 belongs to the Forkhead (Fox) family of TFs, comprising 50 family members in humans which all share the conserved winged-helix domain[122,146]. Because all Fox proteins contain this key structural feature, it is not surprising that other family members in addition to FoxA1, including FoxO1 and FoxE1, have since been characterised as pioneer factors[147,148].

Finally, in addition to the Fox family of pioneer factors, a selected number of other TFs have been shown to have comparatively modest pioneer activity. For example, three of the four Yamanaka factors used to generate iPSCs – Oct4, Sox2 and Klf4 (OSK) - can also recognise partial binding motifs as they exist on nucleosomes[149]. These properties are almost certainly necessary to reprogram differentiated cells with deeply-embedded chromatin profiles into a pluripotent state. The final Yamanaka factor, Myc, binds co-operatively as a conventional TF following prior binding to targets by one or more of OSK. However, despite the emergence of new pioneer factors, their ability to recognise target sites in the context of chromatin appears to be much less efficient than for Fox proteins[149].

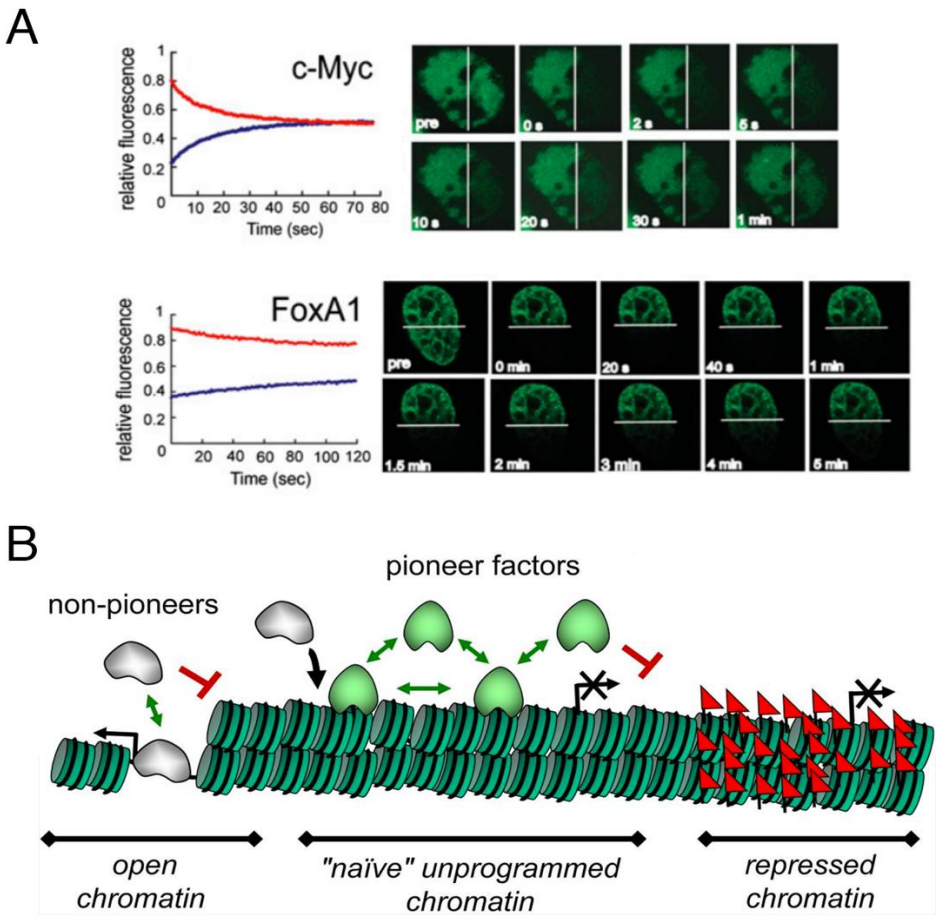


Figure 1:10 **Slow nuclear mobility is a defining feature of the pioneer factor FoxA1.** Non-specific interactions between FoxA1 and histones, via the winged-helix DNA binding domain, give it a greatly reduced nuclear mobility compared to conventional TFs. This is elegantly illustrated in fluorescence recovery after photobleaching (FRAP) experiments using GFP-tagged proteins.

A: While c-Myc, a conventional TF, can return to bleached areas of nuclei in an order of seconds, several minutes pass before a fluorescent signal is observed from FoxA1 returning to bleached areas[145]. (Left) The change in average fluorescence intensities over time are shown for photobleached (blue lines) and unbleached areas (red lines) are plotted, demonstrating a much slower recovery of unbleached FoxA1 molecules to previously bleached areas of nuclei. (Right) Fluorescence microscopy images of a live cell nuclei taken before and after photobleaching demonstrate substantial movement of unbleached c-Myc molecules into bleached regions of nuclei in less than a minute, in contrast to the much slower mobility of unbleached FoxA1 molecules.

B: Schematic model of compacted chromatin scanning by pioneer factors, which is thought to explain slow nuclear mobility of FoxA1[149].

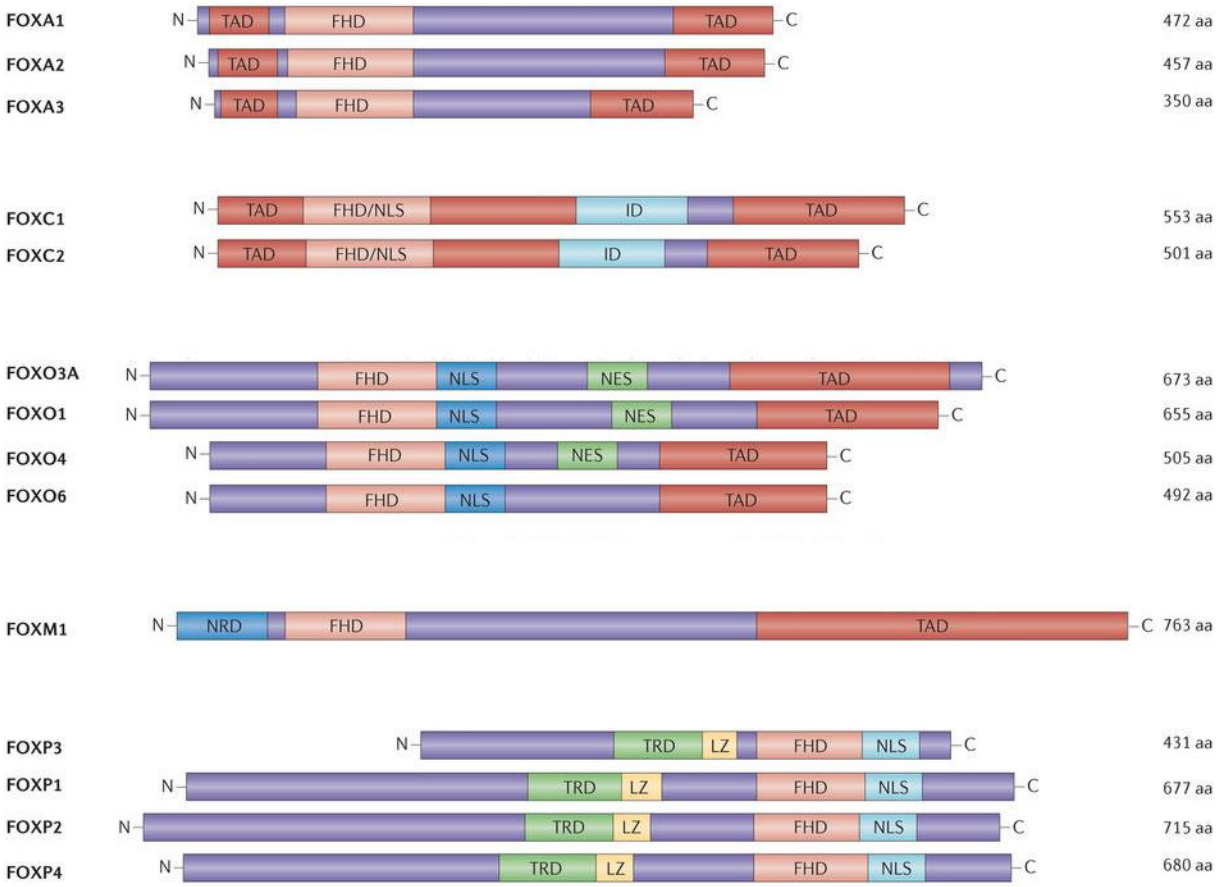


Figure 1:11 **The Forkhead DNA-binding domain is conserved across the FOX protein family.**

As this schematic diagram adapted from Lam *et al.* (2013) demonstrates, different FOX proteins exhibit considerable diversity in the complement of different domains they contain. However, the presence of the winged-helix forkhead DNA binding domain (FHD) that in FoxA1 and FoxO1 confer pioneer activity is a unifying feature across the FOX protein family.

ID = inhibitory domain; LZ = leucine zipper domain; NES = nuclear export signal; NLS = nuclear localisation signal; TAD = *trans*-activation domain; TRD = transcriptional repressor domain.

1.3.7 Non-coding *cis*-regulatory elements and their roles in gene regulation

Just ~1.5% of the human genome sequence encodes protein-coding genes. However, if the remaining DNA did not serve a biological purpose, it is generally considered that it should have been eliminated across generations through evolution, as replication of unnecessary tracts of genome comes at a cost to an organism's biological fitness.

Precisely how much of the remaining non-coding DNA has a biological function is an area of intense debate. Differing accounts ascribe 8 to 80% of the human genome as

functional, depending on the definition and methodology used[151,152]. Disputes notwithstanding, it is widely accepted that a significant portion of the non-coding genome is involved in the regulation of genes through *cis*-regulatory elements. Most genes have several *cis*-regulatory elements, which may be found proximal to the coding sequence, in intragenic regions, or at long distance from the gene they control[153]. Such long-range distal elements can be separated by a few kb to as far as a mega-base (Mb) of intervening DNA sequence, which may contain other genes. The primary roles of *cis*-regulatory elements are to control the timing and levels of target gene expression with TFs in a highly tissue-specific manner. The different classes of these *cis*-regulatory elements shall now be discussed.

1.3.8 Promoters

As previously introduced, promoters are *cis*-regulatory elements that define the TSS of a gene and permit recruitment of the PIC, basal transcriptional machinery and RNA polymerases during gene activation[94]. Despite the commonly accepted functional definition of the core promoter as the minimum DNA sequence required for accurate transcriptional initiation, there is no universal sequence unifying all core promoter elements[154]. For example, the TATA binding element, first identified in archaea and then eukaryotes, was considered for many years to be a ubiquitous promoter element, yet it is only present in 10-20% of human core promoters[155]. The most common sequence feature of many promoters are initiator sequences (Inr) that span the TSSs of both TATA and TATA-less promoters[156,157]. Both TATA box and downstream promoter elements interact with TFIID to promote recruitment of the PIC during transcriptional activation[158]. TATA-less promoters are distinguished by other sequence features, including the GC-box, CCAAT-box and CG islands[154,159]. Unlike TATA-box elements, these elements generally operate independently of sequence orientation and distance relative to the

TSS[159]. Firstly, GC-boxes may contain one or more of the consensus sequence 5'-GGGCGG-3', which are recognised by Sp1 family TFs[160]. The first GC-box promoter was described in the SV40 virus early gene promoter, which is regulated by the ubiquitously-expressed TF, Sp1. GC-box promoters are typically involved in the regulation of housekeeping or ubiquitous genes[161,162]. Secondly, CCAAT-box elements are typically positioned 60-100 bp upstream of the TSS and recognised by NFYA as part of a trimeric complex. The first such elements were identified at the human beta-globin gene promoter[163]. Finally, CG islands are long tracts of typically ~0.5-2 kb of (G+C)-rich sequences that span the TSSs in over 70% of human promoters. Although CG motifs are the normal targets of 5mC deposition by DNMT enzymes, which leads to stable gene repression[36,164], these motifs are predominantly unmethylated in CG islands at constitutive promoters active in most cells.

The organisation of elements at a given gene promoter lies along a continuum between focussed and dispersed[154]. Consistent with their name, focussed promoters have a dense hub of sequence features involved in their regulation, and often appear at tissue-specific genes. In contrast, dispersed promoters have broader domains of elements - CG islands being one example - with most constitutive genes falling under the control of this type of promoter.

1.3.9 Enhancers

Enhancers are the "logic gates" for activation of gene expression, and can greatly amplify the transcriptional activity of the promoters they regulate. Enhancer elements operate in a highly context-specific manner, and therefore play a critical role in establishing tissue-specific gene expression. They are typically ~200-400 bp in size and found distal to their target promoter, most commonly from 1 kb to 1 Mb away. Enhancers feature an increased density of binding motifs for TFs, which bind to increase gene transcription

under the appropriate conditions[165,166]. The first enhancers were described in the early 1980s and included an “upstream modulator” of the sea urchin histone *H2A* gene, the SV40 virus enhancer and the mammalian immunoglobulin gene enhancers[167–172].

Highly regulated genes are typically under the control of multiple enhancers, which either act independently of each other to activate transcription in response to different stimuli or tissue-specific TFs, or synergistically to achieve even greater target gene expression[173,174]. This phenomenon was first demonstrated in elegant studies of the human beta globin locus, which identified five individual elements working in concert as a ‘locus control region’ (LCR) to direct tissue-specific expression. It was also shown that these elements had enhancer activity independently of their relative position to the target gene[175].

While inactive enhancers are normally occupied by nucleosomes, active enhancers acquire increased TF occupancy and an accessible, nucleosome-free chromatin structure detectable as DHS peaks in DNase-seq analyses[2]. Other general features of active enhancers are DNA hypomethylation, enrichment of the histone variants H2A.Z and H3.3, and H3K4me1/2 and H3K27ac modifications.

Finally, enhancers can activate target gene promoters by bringing the two linearly-separated DNA sequences into close physical proximity by DNA looping (Figure 1:12). Such long-range interactions are driven by the enhancer TF-dependent recruitment of the Mediator super-complex, which directly stimulates PIC formation on the target promoter. The assembly of cohesin further stabilises enhancer-promoter interactions[176,177]. Finally, in keeping with physical proximity to the promoter, active enhancers feature bidirectional transcription of enhancer RNAs (eRNAs), perhaps by association with active RNAPII complexes at the promoter during gene activation[178].

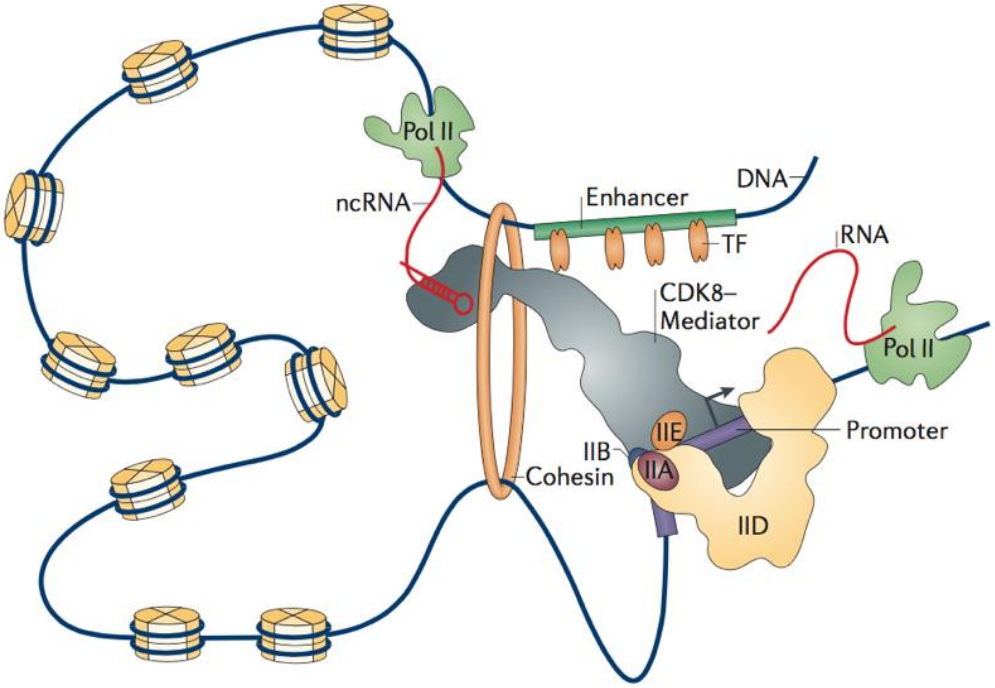


Figure 1:12 **Long-range interactions between distal enhancers and promoter elements.** Three-dimensional chromatin looping interactions are now widely accepted to occur at a substantial proportion of genes during activation. This process is thought to be promoted by TF-dependent recruitment of chromatin architecture proteins, including mediator, which help to bring two sequences that may be separated by megabases of linear genomic distance into close proximity of each other. Such looping interactions are thought to be stabilised by cohesin complexes, allowing enhancer-dependent activation of promoter by currently unclear mechanisms but is thought to involve direct stimulation of the PIC at target gene promoters. Figure from Allen and Taatjes (2015).

1.3.10 Super enhancers

Super enhancers are somewhat arbitrarily distinguished from LCRs as particularly broad domains of individual enhancers that have an especially strong effect on gene activation. They are often described as being an order of magnitude larger in size than LCRs, featuring greater levels of H3K27ac enrichment and Mediator binding, and correlate more strongly with target gene expression than LCRs. Super enhancers commonly regulate critical lineage-determining genes or 'master regulators'[179]. More recently, it has been proposed that super enhancers can also participate in higher order DNA loop structures involving three interacting sequences at key developmental TF genes, such as

the *Sox2* locus in ES cells[180,181]. Collectively, these features are thought to increase the general robustness of target gene activation under the appropriate conditions.

1.3.11 Priming enhancers

Despite large numbers of *cis*-regulatory elements being defined as enhancers according to descriptive features such as enrichment for H3K27ac and DNase I hypersensitivity, up to 80% of these elements lack an ability to directly stimulate gene transcription when directly tested for enhancer function[182,183]. Furthermore, studies from this laboratory identified ~3,000 such elements in T cells that do not typically activate transcription *per se*, but instead maintain an accessible chromatin structure at target genes, priming them for rapidly inducible gene expression[184]. These elements were termed priming enhancers or priming DHSs (pDHSs).

A model example of priming enhancer function is evident at the human *IL3* locus, which is under the control of at least four pDHSs at -1.5 kb, -4.1 kb, -34 kb and -41 kb. These pDHSs did not induce a luciferase reporter driven by the *IL3* promoter in enhancer assays, neither individually nor in combination, in contrast to another classical enhancer at -37 kb which did strongly induce the luciferase reporter[184]. However, CRISPR deletion of the -34 kb pDHS in Jurkat T cells resulted in a diminution of *IL3* accessibility as measured by DNase I hypersensitivity, and a delayed induction of *IL3* expression. In light of this and other evidence, it is likely that a significant portion of *cis*-elements previously defined as classical enhancers will be re-classified in the near future.

1.3.12 Silencers

Opposing the function of the broad class of enhancers, silencers are elements that negatively regulate gene expression. They were originally identified in the 5' UTR of genes, but have since also been found elsewhere. Silencers are the least understood class of *cis*-regulatory element, and are generally thought to work by one of two

mechanisms. Firstly, silencers can act to disrupt assembly of the PIC at the promoter-proximal region of genes, perhaps by recruitment of HDACs, PRC2 complexes, H3K9 methyl transferases or DNMTs that remove histone acetylation and subsequently deposit repressive H3K27me3, H3K4me2/3 or 5mC, respectively. In contrast, silencing elements operating further away from the promoter may contain sites recognised by repressive TFs which disrupt the binding of activating TFs at nearby enhancers[185].

1.3.13 Insulators

Insulators are the final major category of *cis*-regulatory elements, and primarily act to increase the target gene specificity of enhancers. Insulation of gene activity is an important function, since it has been demonstrated that insulator disruption can lead to off-target promoter-enhancer interactions, both experimentally and in cancer[186]. Furthermore, insulator elements can also act as barriers to prevent the inappropriate spreading of repressive heterochromatin into nearby genes[187].

Active mammalian insulator elements are bound by the protein CTCF [188]. CTCF interacts with Cohesin and these complexes mediate between insulator elements to bring linearly-separated DNA sequences into close physical proximity via DNA looping, establishing segregated functional domains[189] referred to as 'insulated neighbourhoods' . Such interactions can be short range, partitioning the activity of one gene from another a few kb away, or they can be separated by many hundreds of kb of intervening DNA sequence.

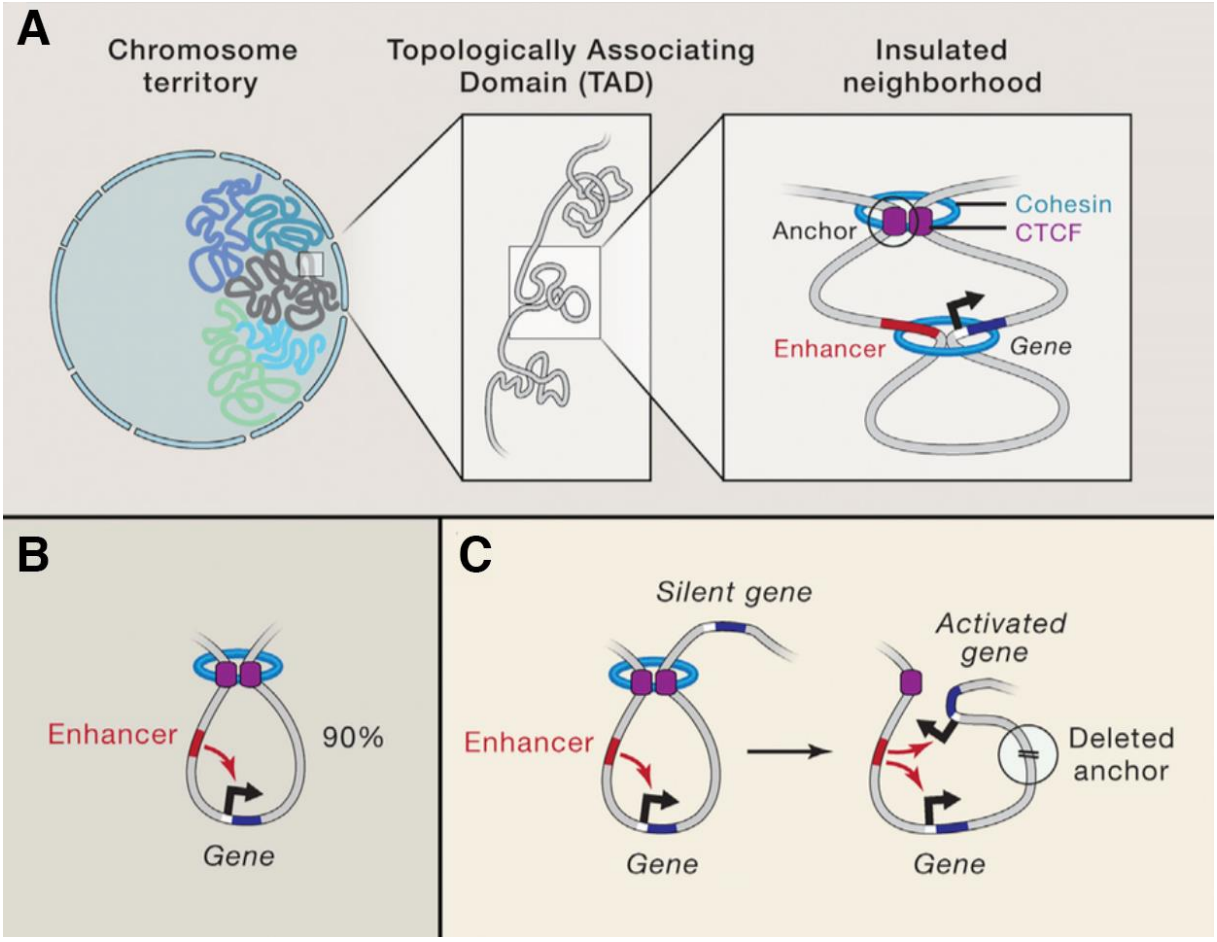


Figure 1:13 **Insulated neighbourhoods partition active genes from silent ones.**

A: Further detail of the organisational hierarchy of chromosomes: the TADs previously described in 1.1.3 are partitioned further into insulated neighbourhoods.

B: stereotyped example of an insulated neighbourhood, where a gene and its cognate enhancers are physically partitioned by cohesin and CTCF to prevent the activation of other nearby genes.

C: deletion or disruption of CTCF sites that participate in gene insulation can lead to the ectopic reactivation of other local genes.

Adapted from Hnisz *et al.* (2016).

Further increasing the specificity of insulator function, these CTCF sites must be oriented towards each other for dimerisation and loop formation to occur[190]. The partitioning of the genome into insulated neighbourhoods was demonstrated to be functionally significant in human ES cells, where over 90% of all detected promoter-enhancer interactions occurred within the boundaries of insulated neighbourhoods[191]. Insulated neighbourhoods are also referred to as sub-TADs, a constituent of the TADs described in

1.1.3, and are generally constitutive across different cell types[191]. In some instances, insulators have also been defined as silencers due to their enhancer-blocking activity. In the case of the lysozyme gene, activation of non-coding transcription from an enhancer was sufficient to suppress a silencer by removing CTCF from its binding site[192,193].

Finally, CTCF binding at some specific sites is highly sensitive to DNA methylation, and cancers with aberrant DNA methylation function often feature abnormal insulator or TAD patterns resulting from disrupted CTCF binding. Insulator dysfunction can also lead to the ectopic activation of oncogenes[194].

1.4 Haematopoiesis and the haematopoietic stem cell

Haematopoiesis is the process by which all blood cell types are produced. Despite their vast functional diversity, mature blood cells arise from a common stem cell with the potential to propagate all lineages. The existence of such cells was first postulated in cytological studies by Maxinow in the 1900s, who observed a "ubiquitous, non-differentiated" cell type in foetal blood that could "produce a variety of differentiation products"[195].

However, it was not until the 1960s that a functional understanding of haematopoiesis began to emerge with the work of Till and McCulloch, who demonstrated that rare cells from the bone marrow could clonally repopulate *in vivo*, forming macroscopic colonies on recipient mice spleens[196,197].

With the development of fluorescence-activated cell sorting (FACS), it became possible to separate cells by their immunophenotype, leading to the isolation of what we know today as haematopoietic stem cells (HSCs) by Weissman and colleagues at Stanford[198]. These HSCs were operationally defined by their ability to fully reconstitute the haematopoietic system in lethally irradiated recipient mice[199]. HSCs have an innate

capacity to self-renew, ensuring the production of comparatively short-lived mature blood cells is sustained throughout life. Thanks to sustained intense research, the haematopoietic system is perhaps the best understood adult developmental pathway in the body and HSCs have become a paradigm for the general study of stem cell biology in other tissues.

1.4.1 Discovery of the haematopoietic stem cell niche

It was known for many years from work by Till and McCulloch that adult HSCs reside in the bone marrow, and that this environment was somehow critical to their function[196]. However, HSCs are extremely rare, accounting for ~0.003% of all cells in the bone marrow[200]. HSCs occupy one of two microenvironments in the bone marrow, called the endosteal and perivascular niches. These anatomically distinct niches also have functional implications for the HSCs embedded within, and an understanding of their highly complex nature only recently started to become clear.

The endosteal niche is located towards the outer edge of the bone marrow, is hypoxic, and is enriched for quiescent long term HSCs (LT-HSCs). It is not fully clear why LT-HSCs preferentially localise to this niche, but paracrine signalling from nearby osteoblasts is thought to help maintain LT-HSC quiescence[201]. Hypoxia is also likely to promote the long-term potential of LT-HSCs by enforcing metabolic dormancy and minimising mutations arising from oxidative stress[202,203].

By contrast, the perivascular niche at the core of the bone marrow is distinguished by extensive sinusoidal vasculature, is highly oxygenated, and occupied by actively cycling short term HSCs (ST-HSCs). Perivascular mesenchymal stem cells (MSCs) secrete cytokines critical to the maintenance of ST-HSCs, including CXCL2 and stem cell factor (SCF)[204]. There is evidence from mice that vascular endothelial cells also support ST-

HSCs via VEGFR2 signalling[205]. Furthermore, ST-HSC mobilisation to peripheral blood appears to be subject to circadian regulation by sympathetic nerve fibres acting on perivascular cells[206,207]. With appropriate stimuli, ST-HSCs divide asymmetrically to give rise to the progenitors that go on to populate all differentiated cell types of blood via haematopoiesis.

1.4.2 The classical model of haematopoiesis

In the years following the isolation of HSCs in 1986, further work pioneered by the Weissman laboratory characterised phenotypically discrete subsets of precursor cells, suggesting that adult haematopoiesis is a hierarchical process[198,199,208,209]. At the apex of this hierarchy are multipotent, self-renewing HSCs which have the capacity to give rise to all mature blood cell types via ordered fate restriction events involving progressively less potent progenitor cells[210] (Figure 1:14).

The HSC compartment is divided into two subsets of cells according to their capacity for self-renewal. The more 'primitive' LT-HSCs are the most quiescent and capable of self-renewal throughout adult life. By contrast, ST-HSCs retain multipotency but are more actively cycling than their long-term counterparts and have a comparatively lower self-renewal capacity[211].

By the time cells enter the multipotent progenitor (MPP) compartment, self-renewal is lost entirely in favour of greater cell proliferation. Crucially, although these cells retain sufficient potency to differentiate into any mature lineage, MPPs are already developmentally committed and cannot re-enter the HSC compartment[209].

The first lineage bifurcation in this model occurs following the MPP stage, where cells become committed to either lymphoid or myeloid/erythroid lineages. These are termed the common lymphoid progenitor (CLP) and common myeloid progenitor (CMP)

compartments, respectively. CLPs give rise to all lymphoid cells, including B cells, T cells, NK cells and others comprising the adaptive immune system. Lymphoid development occurs in the bone marrow until immature lymphocytes emerge, which enter the circulation to mature at distal sites including the thymus (T cells) and germinal centres (B cells)[212,213].

In contrast to CLPs, CMPs in this model retain potency to generate either myeloid or erythroid cells following progression into one of two more specified compartments: the megakaryocyte/erythroid progenitors (MEPs) and granulocyte/monocyte progenitors (GMPs)[208]. GMPs ultimately give rise to all mature myeloid lineages in this model.

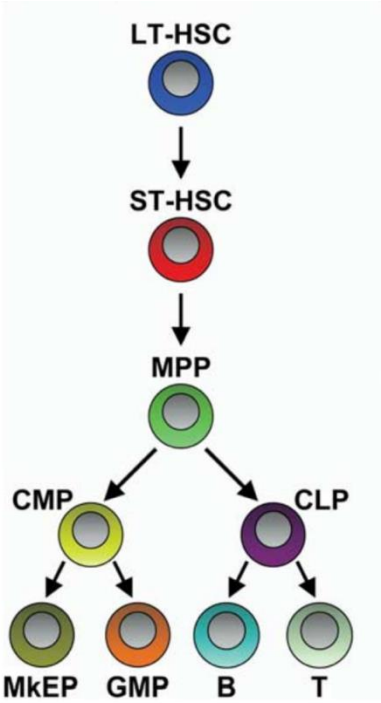


Figure 1:14 **The classical Weismann model of haematopoiesis.**

In the Weismann model, an HSC undergoes defined step-wise transitions during haematopoiesis. The defining lineage commitment step in this model occurs at the MPP state, where cells are thought to commit strictly to either the myeloid/erythroid or lymphoid lineages, via a CMP or CLP progenitor state respectively. Adapted from Adolfsson *et al.* (2005).

LT-HSC = Long-term haematopoietic stem cell; ST-HSC = short-term haematopoietic stem cell; MPP = multipotent progenitor; CMP = common myeloid progenitor; CLP = common lymphoid progenitor; GMP = granulocyte/monocyte progenitor; MkEP = megakaryocyte/erythroid progenitor.

1.4.3 Updated models of haematopoiesis

Since this classical model of haematopoiesis was proposed, more recent observations from the Jacobsen laboratory have prompted revision of key aspects of the model. Firstly, an additional compartment of progenitors was identified, featuring the expression of lymphoid marker genes and FMS-like tyrosine kinase (Flt3). These cells could generate both lymphoid and myeloid cells, but not erythroid lineages, and thus were named lymphoid-primed multipotent progenitors (LMPPs)[214]. Secondly, it was observed that some MEPs were derived directly from MPPs without passing through an intermediate CMP state, suggesting that the strict lymphoid or myeloid partitioning is not the earliest lineage restriction step[215]. Taken together, the accumulated evidence support a revised model of haematopoiesis featuring a greater degree of plasticity in fate restriction events, which are perhaps determined in a probabilistic manner by heterogeneity within the ST-HSC and MPP compartments brought about by both intrinstic and extrinsic factors (Figure 1:15).

Despite greatly expanding our understanding of how blood cells are produced, two main limitations of historic approaches to studying haematopoiesis have become apparent. Firstly, the model was largely built up from studies of cell populations purified using subjectively-defined FACS sorting gates and immunophenotypes. Bulk analyses of sorted populations are reliant on average molecular profiles, masking functionally significant heterogeneity between individual cells. Furthermore, although single-cell analyses on these populations may reveal underlying patterns of sub-population diversity, transitory changes occurring across FACS sort gates are impossible to study[208,216].

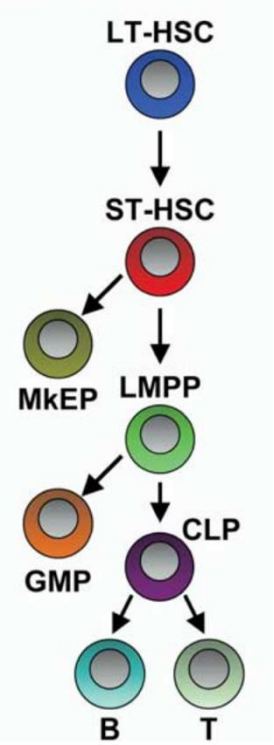


Figure 1:15 **The updated Jacobsen model of haematopoiesis**.

In contrast to the Weissman model, the Jacobsen model proposes that some lineage restriction takes place before the 'MPP' state, in the ST-HSC compartment. ST-HSCs may directly commit to the erythroid/megakaryocyte lineages at this stage, or instead develop into lymphoid-primed multipotent progenitors (LMPPs) that can generate myeloid and lymphoid, but not erythroid lineages. Adapted from Adolfsson *et al.* (2011).

Secondly, there is longstanding concern over the reliance on colony forming assays and transplantation as a reliable proxy for the in vivo, homeostatic haematopoietic environment. This issue is underlined by renewed debate over where lineage bifurcations occur, partly arising from the FACS marker and experimental assay used[215,217].

Significantly, the emergence of methods to profile chromatin and gene expression in single cells is prompting a complete reconsideration of the classical model of haematopoiesis. Compelling data from recent studies suggest that lineage commitment in fact starts at a much earlier stage than previously thought: in the phenotypic HSC (CD150+CD34-Flt3-KSL) compartment. Traceable barcoded cells from this population

have been shown to directly give rise to unipotent progenitors without need for an oligopotent intermediate[218,219], consistent with earlier reports from the Jacobsen laboratory regarding the origin of MEPs[215].

A more recent investigation using index-sorted single cell HSPCs, with parallel measurement of immunophenotype and either transcriptome or functional output in the same cell, suggested that previously defined compartments such as MPPs and LMPPs are not discrete cell types *per se*, but rather transitory states that exist along a continuum of "cloud-HSPCs", with a higher probability for a single lineage commitment[220]. These results also suggested that in adult haematopoiesis, oligopotent lineage priming is a comparatively rare event, with the overwhelming majority of cells being either multipotent or already committed to a single lineage. These findings corroborate fate-mapping experiments of endogenously-labelled HSCs in mice, where immunophenotypically-equivalent stem cells were found to have diverse enhancer DNA methylation patterns priming distinct lineage outputs[221].

The most recent model of haematopoiesis considers the latest insights into lineage commitment as shown in (Figure 1:16). As the power and robustness of single cell approaches continue to develop at pace, it is likely that our understanding of haematopoiesis will be transformed further in the immediate future.

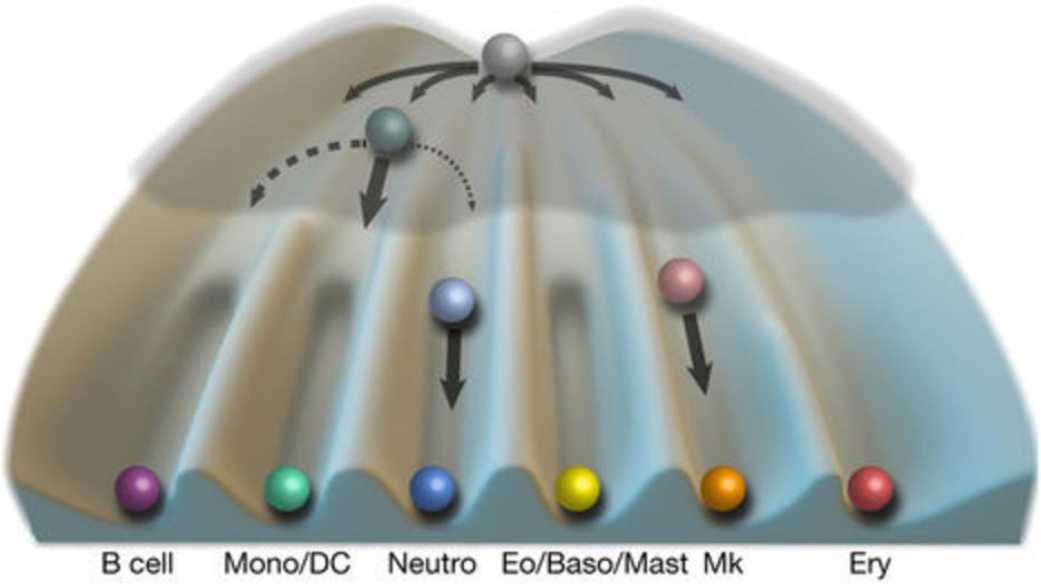


Figure 1:16 **The cloud-HSPC model of haematopoiesis.**

The most recently proposed model of haematopoiesis postulates there are no clearly-defined MPP or LMPP compartments. Instead, HSCs have at least some degree of predetermined lineage bias, which may be overcome in response to the balance of external cues. Considerable epigenetic heterogeneity within the HSC compartment is partly thought to underpin this model. Figure from Velten *et al.* (2017).

1.4.4 Transcription factors regulate myelopoiesis

Lineage specification of HSPCs into mature blood cells is associated with a loss of self-renewal and the parallel acquisition of mature cell characteristics. These outcomes result from dramatic shifts in gene expression; the repression of self-renewal genes and concurrent activation of differentiation-specific genes. Genes involved in these processes exist as nodes in highly-connected networks, where individual genes may antagonise or synergise with one another following the tightly-regulated activities of lineage-specific TFs. Of all the pathways involved in haematopoiesis, the instrumental roles of TFs are most well known in myeloid cell production, or myelopoiesis (Figure 1:17).

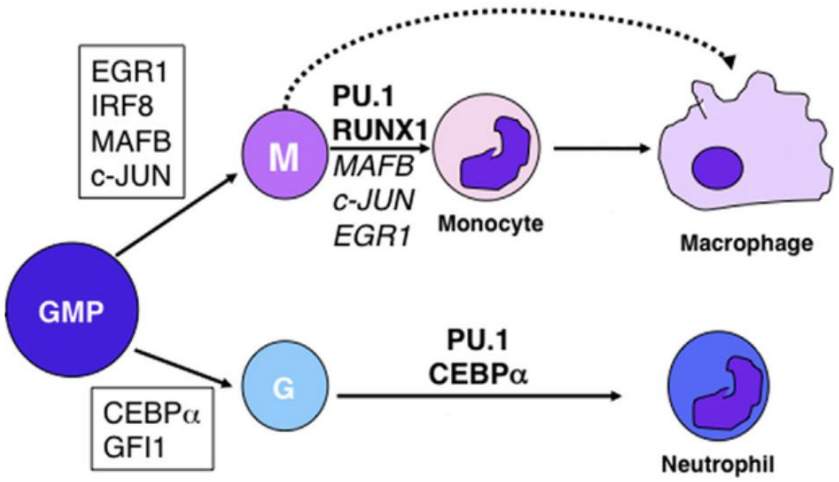


Figure 1:17 **Defined roles for transcription factors in myelopoiesis.**

The most critical regulators of myelopoiesis are PU.1 and C/EBPα, which function in a highly dose-dependent manner to regulate lineage commitment from the GMP stage to either granulopoietic (G) or monopoietic (M) progenitors. These factors operate in gene regulatory networks with other TFs characterised by more restricted expression patterns to promote terminal differentiation.

Atop the regulatory hierarchy in myelopoiesis is the transcription factor RUNX1. RUNX1 contains a conserved 'Runt' DNA-binding domain which recognises 5'-YGYGGTY-3' sequences to regulate the expression of target genes[222]. RUNX1 is actually a master regulator throughout haematopoiesis, required for the emergence of definitive haematopoietic cells from haemogenic endothelium during embryogenesis and expressed across adult haematopoiesis[223–225]. In myelopoiesis, RUNX1 promotes the expression of key myeloid genes including *IL*-3, *CSF1R* and - critically – the myeloid TF genes *SPI1* and *CEBPA*[226,227].

The first stage of myelopoiesis occurs when MPPs/LMPPs commit to the myeloid pathway via the CMP state instead of entering the lymphoid pathway. Two key features of this event are the continued expression of PU.1 (*SPI1*), a member of the ETS family of TFs, and the parallel downregulation of GATA1. The role of PU.1 in myelopoiesis was first illustrated using reverse genetic approaches: *Spi1*-/- mice are embryonic lethal, likely due to impaired erythroid differentiation, but also lack myeloid differentiation[228]. More

recent, conditional knockout studies showed that *Spi1*-/- HSPCs failed to fully reconstitute the haematopoietic system, including defects again at the CMP and CLP level[229]. In the normal context, PU.1 activates myeloid-specific target genes[230].

A requirement for downregulation of GATA1 during CMP lineage commitment was demonstrated in vitro, where forced expression of GATA1 blocked myelopoiesis and expanded the erythroid compartment[231,232]. The antagonistic relationship between GATA1 and PU.1 emerged with co-immunoprecipitation assays showing that GATA1 binds to the PU.1 DNA binding domain, inhibiting its ability to activate pro-myeloid genes[233]. Further chromatin immunoprecipitation experiments showed that GATA1 and GATA2 bind to *SPI1* regulatory elements and act sequentially to repress its expression[234]. The assumption that PU.1 and GATA1 expression levels determine myeloid lineage commitment was recently disputed. Continuous analyses of protein expression levels for these TFs in single cells suggested that early myeloid restriction is not determined simply by relative levels of PU.1 and GATA1. Moreover, the role of these TFs is to reinforce fate decisions determined by other mechanisms[235].

The next fate restriction occurs at the CMP/GMP transition. After this step, cells are capable of generating myeloid lineage cells only. Progression to the GMP compartment requires the expression of C/EBPα, a TF containing a common bZIP DNA-binding domain shared with all six human C/EBP family members[236,237]. This requirement was demonstrated in Cebpa-/- mice, which lack GMPs and granulocytic lineages, but not monocytes[238]. However, this seems to be a stage-specific requirement, as conditional deletion of C/EBPa in mouse GMPs was not found to disrupt in vitro granulopoiesis[239]. Like PU.1, C/EBPα is known to activate myeloid-specific genes, including M-CSF[240].

Following the GMP stage, myeloid precursors commit to either the monocyte or granulocyte lineages. The balance between PU.1 and C/EBPα expression levels is an important feature of this process. In GMPs, moderate levels of PU.1 regulate a bipotent network of genes involved in both monocyte and granulocyte development. However, increasing levels of PU.1 shifts the GMP gene expression programme towards the monocyte/macrophage lineage fate, in a regulatory circuit involving IRF8 and the AP-1 family of TFs[241]. IRF8 co-operates with PU.1 to activate expression of NF1, a feature of differentiating monocytes[242]. AP-1 proteins, including c-Jun and c-Fos, work alone and with other TFs to reinforce the monopoietic expression programme. For example, PU.1:c-Jun complexes induce the expression of Interleukin-1β and other monocyte/macrophage-restricted genes greater than PU.1 alone[243]. In contrast, C/EBPα and c-Jun can interact via their leucine zipper domains, inhibiting the ability of C/EBPα to bind to DNA and activate granulocytic target genes[244]. Taken as a whole, the effects of this subset of TFs re-inforce the monocyte/macrophage lineage commitment.

Conversely, granulocyte lineage commitment involves the upregulation of *CEBPA* and granulopoietic factors, including the transcriptional repressor *GFI1*[245]. GFI1 participates in a feed-forward mechanism in granulopoiesis by repressing PU.1-dependent activation of monopoietic genes[246]. The final specification of granulocytes into terminally differentiated neutrophils and eosinophils requires the continued activity of GFI1 to reinforce repression of immature target genes, and the upregulation of C/EBPε which activates more mature target genes[247].

1.5 Acute Myeloid Leukaemia

Acute Myeloid Leukaemia (AML) is a clonal haematological malignancy characterised by the rapid, uncontrolled expansion of abnormal myeloblast cells. If these blasts exceed 20% of bone marrow cellularity, a clinical diagnosis of AML is made. The incidence of

AML is most common in adults, and accounts for 25% of all adult leukaemias. The rapid expansion of abnormal blasts interferes with normal haematopoiesis and leads to leukaemic cell infiltration of peripheral organs including the spleen, liver, lungs and brain. Without treatment, AML generally progresses rapidly, causing bone marrow failure, infection and fatality in a matter of weeks[248].

AMLs are classified into several subtypes according to one of two commonly used systems; the French-American-British (FAB) system or the World Health Organisation (WHO) system. The FAB system classifies AMLs on the basis of cell morphology and the level of leukaemic cell differentiation, while the more recently introduced WHO system adopts more specific criteria including cytogenetic risk and specific genetic aberrations[249].

1.5.1 Mutations co-operate to drive clonal transformation in AML

AML, like many neoplasms, has long been considered to be a clonal malignancy arising from a single cell of origin[250]. In this model, the accumulation of oncogenic mutations in 'founder' cells permit the expansion of multiple clones which compete with one another for finite resources according to Darwinian selection[251]. Eventually, a clone with advantageous mutations emerges as the dominant population detected in bulk AML samples at presentation. Gilliland and Griffin first proposed that two classes of mutations are necessary for AML to develop in their so-called 'two-hit' model[252]. Mutations were categorised into two classes according to their functional consequences, but at least one from each class is required for leukaemogenesis. Class I mutations confer enhanced proliferation, whereas class II mutations suppress normal myeloid differentiation.

Consistent with the property of increasing cell proliferation, class I mutations most commonly affect mitogenic signalling molecules. A good example of a class I mutation is

the internal tandem duplication (ITD) of *FLT3*. The wild-type *FLT3* gene encodes a receptor tyrosine kinase which is activated upon extracellular binding of FLT3 ligand (FL)[253]. In normal cells, FLT3 signalling stimulates proliferation of ST-HSCs and, to a lesser extent, early B cell progenitors[254,255]. This process is tightly regulated, not only by control of FL secretion, but also via the juxtamembrane (JM) domain of FLT3 itself, which mediates auto-repression of signalling after activation by FL. However, *FLT3*-ITD mutations generate in-frame insertions within the JM domain from 3 to over 400 bp in size, nullifying this auto-repressive activity[256]. The result is constitutive ligand-independent activation of FLT3-ITD signalling and ectopic activation of wild-type FLT3 in *trans*[257]. The ultimate consequence of this is sustained proliferative signalling, driving the survival and growth of affected cells[252].

Class II mutations slow or block normal myeloid differentiation, and often directly target TF genes. The mechanisms through which these mutations act include the inappropriate re-targeting of mutated TFs to oncogenic loci, or by disrupting the ability of mutated TFs to induce the expression of differentiation-specific genes that would normally be activated by the wild-type allele. The two best examples of class II mutations in AML affect the master regulator TF genes *RUNX1* and CEBPA, and are explored in detail in 1.5.4 and 1.5.5.

The two-hit model was informative in developing an initial understanding of acute leukaemogenesis, but was also too simplistic to categorise certain other mutations which did not exert strict binary effects on either proliferation or differentiation. However, these other mutations shared a common effect of deregulating the epigenome to aberrantly maintain self-renewal. Thus, a third class of AML mutation was devised for epigenetic regulators[258]. Class III mutations act via diverse mechanisms, but commonly involve deregulation of DNA methylation or histone modification processes, leading to ectopic

repression of tumour suppressors and/or activation of oncogenes. Such mutations may overtly affect such regulators – as in *DNMT3A* mutations which directly impair DNA methylation - or they may deregulate epigenetic processes via indirect mechanisms. The best examples of indirect class III mutations are those that affect the isocitrate dehydrogenase enzyme genes *IDH1* and *IDH2*, occurring in up to 20% of *de novo* AMLs[259]. These enzymes are essential for mitochondrial oxidative phosphorylation, catalysing the decarboxylation of the TCA cycle intermediate isocitrate into α-Ketoglutarate. However, mutation to either of these genes occurs in up to 20% of *de novo* AMLs and causes dominant negative inhibition of wild-type *IDH1* and *IDH2,* and accumulation of the oncometabolite D-2-hydroxyglutarate (2HG). 2HG accumulation leads to hypermethylation of 5mC across the genome and the parallel inhibition of TET2-mediated hydroxymethylation, resulting in the global repression of myeloid-specific genes, particularly those driven by CpG promoters.

The most common examples of the three mutation classes are summarised in

Table 1**:2**.

Table 1:2 Commonly occurring mutations in AML and their consequences.

Mutations in AML commonly alter the function of signalling molecules, transcription factors or epigenetic regulators. However, many frequently occurring mutations do not fall strictly into one category; *NPM1* and *MLL* mutations are notable examples. The maximum reported clinical frequencies for each mutation in *de novo* AMLs are indicated in brackets[251,258,260].

Class I mutations **Signal transduction** Proliferative advantage	**Class II mutations** **Transcription factors** Differentiation block	**Class III mutations** **Epigenetic regulators** Diverse Consequences
(30%) *MLL* rearrangement	*MLL* rearrangement	*MLL* rearrangement
(21%) *NPM1*	*NPM1*	*NPM1*
(22%) *RUNX1* point mutation	*RUNX1* point mutation	(23%) *TET2*
(25%) FLT3-ITD	(20%) C/EBPα pt. mutation	(33%) *IDH*
(4%) FLT3-TKD	(5%) CBFβ/MYH11	(22%) *DNMT3A*

(10%) *NRAS*	(10%) RUNX1/ETO	(5%) *ASXL2*
(2%) *KIT*		(N.D.%) Cohesin family

1.5.2 Clonal evolution of AML most commonly begins with pre-leukaemic mutations in the HSC compartment

Despite a more detailed knowledge of the types of genetic abnormalities present in AML, the question of which cell population sustained the first pre-leukaemic mutation remained unresolved. Even prior to the identification of class III mutations, it was presumed that the sheer numbers of leukaemic blasts produced in AML must require self-renewal, a property lost beyond the ST-HSC compartment in normal haematopoiesis[209]. Thus, it was suggested that the first mutational event in AML occurred in HSCs. At the time, this proposal seemed inconsistent with existing evidence that HSCs were protected against mutation, but more recently it has been demonstrated that quiescent HSCs do accumulate potentially deleterious DNA lesions during aging that could promote the initial propagation of pre-leukaemic clones[261].

The first compelling evidence that leukaemogenic transformation begins in the HSCs emerged from studies by the Dick laboratory, which demonstrated that leukaemia initiating cells from human AML patients bore the same immunophenotype of normal blood-reconstituting cells. These cells, termed SCID leukaemia initiating cells (SL-ICs) for their ability to engraft leukaemias in SCID xenograft mice, featured a common $CD34^{+}CD38^{-}$ primitive normal cell immunophenotype, regardless of the subtype of AML from which they were derived. Critically, serially transplanted SL-ICs differentiated *in vivo* to leukaemic blasts, suggesting that the development of AML is a hierarchical process not unlike normal haematopoiesis[262].

A more recent single cell analysis of *de novo* AMLs revealed that genetically-distinct leukaemic sub-clones featured common ‘founder’ mutations already present in pre-

leukaemic HSCs[263]. These founder mutations commonly affected epigenetic regulators, abnormally maintaining self-renewal beyond the HSC compartment. Further mutations required for complete transformation were acquired in a stepwise manner, giving rise to a heterogeneous landscape of individual leukaemic clones sustained by diverse abnormalities.

The clonal evolution model of AML development has important implications for therapy. Firstly, it explains why many AML patients may initially respond to chemotherapy but then develop resistant disease. Primary treatment may effectively target the dominant population of therapy-sensitive leukaemic clones, but in the process imposes a selective pressure on residual clones that can promote further transformation. Relapse can be driven either by pre-existing therapy-resistant clones emerging as the new dominant clone, or by further mutation of therapy-sensitive cells to acquire resistance[264]. These two mechanisms of relapse were both confirmed to occur in a very recent study comparing paired presentation/relapse AML samples[265]. Above all, the clonal evolution model highlights a need to better understand the underlying clonal composition of leukaemias at presentation, as bulk analyses may not be sensitive enough to detect minor sub-clones that act as a reservoir for relapse following primary treatment.

Although it is now thought that most *de novo* AMLs form from pre-leukaemic clones derived from the HSC compartment, exceptions do exist. Perhaps the best example is Acute Promyelocytic Leukaemia (APL), a subtype of AML driven by the *PML-RARA* fusion gene. Despite *PML-RARA* being the primary leukaemia-initiating mutation in APL, in one study the fusion transcript was detected only in the more mature CD34+CD38+ compartment, suggesting that these cells became transformed at a much later stage in myeloid differentiation than other subtypes of AML[266]. There is still evidence that pre-leukaemic APL cells evolve clonally like other AMLs, except that the 'founder clone' is

already at a more differentiated state than an HSC. This exceptional characteristic of APL likely accounts for the unusually favourable prognosis compared to other AMLs[267].

Whilst a mature cell of origin in APL represents a comparatively rare event in acute myeloid leukaemias, other haematological malignancies commonly arise from more differentiated founder cells. For example, perhaps half of chronic lymphocytic leukaemia (CLL) cases are thought to originate from mature B cell precursors which have already undergone somatic hypermutation[268]. Furthermore, cases of both acute B and T lymphoblastic leukaemias have been reported to arise from cells that have already left the HSC compartment[269,270].

1.5.3 The development of AML involves deregulation of critical haematopoietic transcription factors

As described in 1.4.4, myelopoiesis is tightly regulated by the stage-specific expression and activity of TFs, underlining their importance in normal myeloid development. A common feature of cellular transformation in AML is the disruption of these and other TFs, resulting in the ectopic activation of proliferation and self-renewal genes, and the repression of terminal myeloid differentiation genes. Deregulation of TF function can either result directly from mutation in the case of class II lesions which affect TF genes themselves, or it can be an indirect consequence of signalling or epigenetic regulator mutations.

Although class I mutations affect signalling proteins, the downstream targets of deregulated signalling pathways in AML are often themselves TFs. An excellent example of this is the found in AMLs with *FLT3* ITD mutations. Chronic FLT3-ITD signalling leads to hyperactivation of downstream signalling pathways, predominantly MAP kinase, causing inappropriate activation of TFs including AP-1 proteins and RUNX1 and the induction of leukaemogenic target genes[271].

Class III mutations also indirectly affect TF function. For example, genome-wide DNA hypermethylation in AMLs with *IDH* mutations disrupts the binding of GATA factors and EVI1 to target myeloid-specific genes, suppressing normal myeloid differentiation[259]. In the context of glioma, ectopic genome methylation resulting from *IDH* mutations abrogates the recruitment of CTCF to critical insulator sequences and causes to the ectopic activation of *PDGFRA*[194] (Figure 1:18). Although such specific examples of insulator dysfunction have not been described in the context of AML, it is almost certain that these are recurring phenomena that occur in leukaemic transformation.

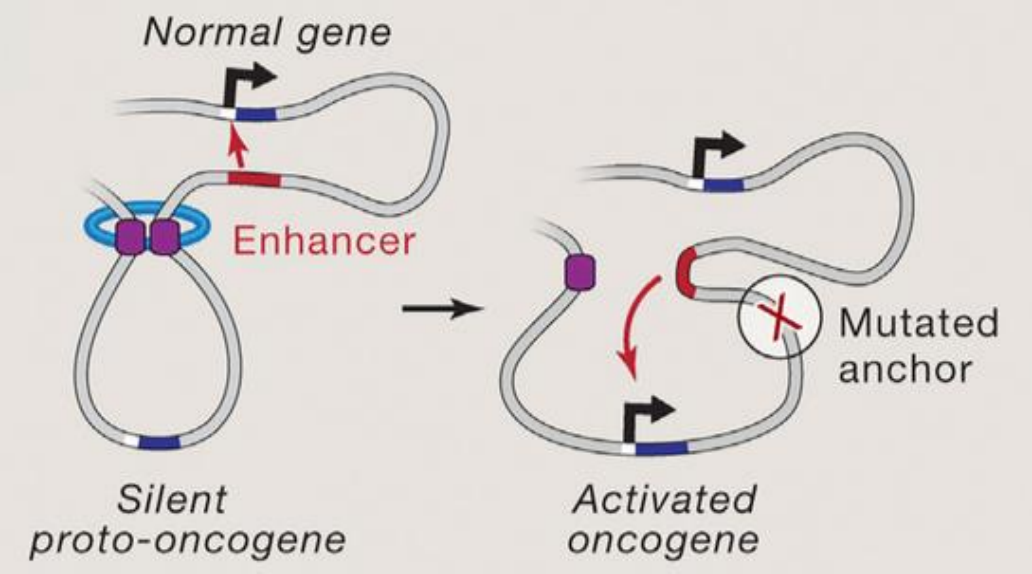


Figure 1:18 **Disruption of insulators can promote the ectopic activation of oncogenes.**

Processes which perturb normal insulator function, including hypermethylation/mutation of CTCF binding sites, can ablate insulated neighbourhoods partitioning silent oncogenes from the nearby active genes, leading to ectopic expression. Adapted from Hnisz *et al.* (2016).

1.5.4 Disrupted RUNX1 activity is a recurrent feature in leukaemogenesis

Given the critical roles for RUNX1 in myelopoiesis described in 1.4.4, it is unsurprising that RUNX1 is among the most commonly mutated TFs in AML, occurring in ~10% of all cases[272,273]. Generally speaking, mutations directly affecting RUNX1 function are class II,

presenting a differentiation block[274]. However, deregulation of wild-type RUNX1 function also frequently results from class I and class III mutations[258,271].

The most common cytogenetic abnormalities in AML affect *RUNX1* or its essential co-factor *CBFB*, generating a range of chimeric proteins including RUNX1-ETO, RUNX1-EVI1 and CBFß-SMMHC (Figure 1:19). The t(8;21) chromosomal translocation creates RUNX1-ETO, fusing the Runt domain of RUNX1 to the near-complete ETO polypeptide[275], with several consequences that block myeloid differentiation. Firstly, the repressive functions of wild-type ETO are combined with the DNA-binding activity of RUNX1, resulting in the targeted recruitment of the NCOR1/HDAC co-repressor complex to myeloid genes usually activated by wild-type RUNX1, including *CEBPA*[276–278]. Secondly, recent genome-wide surveys of RUNX1-ETO binding in AML demonstrated that the RUNX1-ETO outcompetes wild-type RUNX1 activity at pre-existing open chromatin sites[279]. An important contributor to this competitive binding is the ability of RUNX1-ETO to multimerise at target loci, stabilising a higher order complex containing the fusion oncoprotein[280]. RUNX1-ETO also directly inhibits the myeloid regulators PU.1 and C/EBPα via protein-protein interactions, suppressing the activation of differentiation-specific target genes[281]. The profound effects of RUNX1-ETO in AML were demonstrated to be reversible following targeted knockdown of the fusion protein, resulting in widespread epigenetic reprogramming, loss of self-renewal and induction of myeloid differentiation. This was largely due to restoration of normal PU.1 and C/EBPα activity[282].

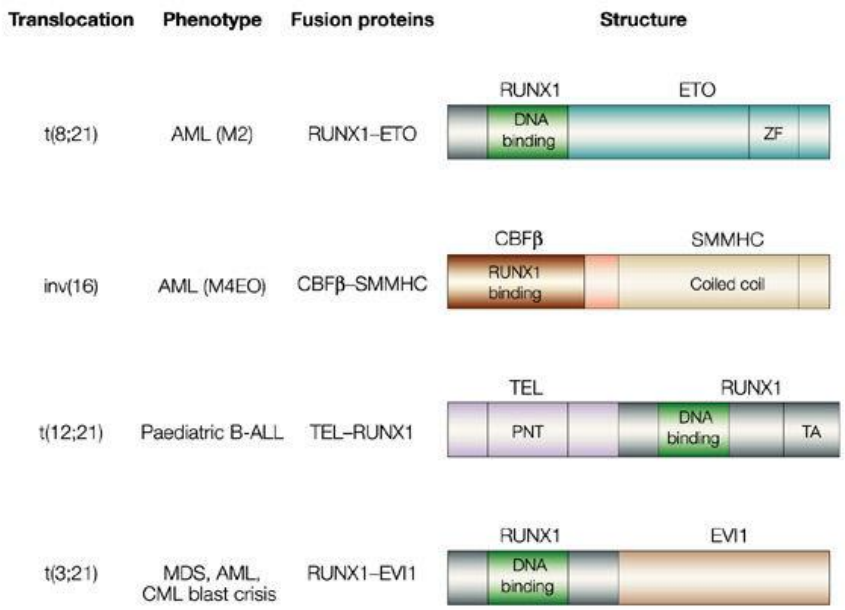


Figure 1:19 **RUNX1 and its binding partner, CBFβ, are frequently rearranged in AML.**

Over a quarter of adult AML cases are characterised by translocations that disrupt the function of RUNX1 through diverse mechanisms. Figure from Speck and Gilliland (2002).

Another genomic rearrangement affecting RUNX1 in AML is the t(3;21) translocation, fusing the Runt domain of RUNX1 to the entire EVI1 protein to generate the RUNX1-EVI1 oncoprotein[283]. EVI1 contains two zinc-finger DNA-binding domains and a conserved SET domain with histone methyltransferase activity[284]. Knockout experiments have demonstrated that wild-type *Evi1* is an essential positive regulator of self-renewal in HSCs[285]. Similar to RUNX1-ETO, the main consequence of RUNX1-EVI1 is a block in myeloid differentiation via downregulation of myeloid genes including *CTSG, CSF1R* and *CEBPA*. This is accompanied by the parallel upregulation of self-renewal genes including *HOXA9* and *MEIS1* via ectopic recruitment of GATA2 by the EVI1 portion of the oncoprotein.

Other chromosomal abnormalities disrupting normal RUNX1 activity include the chromosome 16 inversion, which fuses the RUNX1-binding domain of CBFß to the coiled-coil domain of SMMHC, generating the *CBFB-MYH11* fusion gene. The

oncoprotein product exerts a dominant negative function over wild-type CBFß by disrupting its ability to associate with RUNX1 and stabilise its DNA-binding activity[286,287].

Critically, the reliance of AML cells on these diverse oncoproteins make them attractive candidates for targeted drug therapies. Structure-based design studies led by the Bushweller group aim to address these opportunities to improve patient outcomes in the clinic[288].

In addition to being deregulated through chromosomal rearrangements, hetero- and homozygous point mutations to *RUNX1* can also lead to the development of AML through similar downstream effects[289]. Finally, roles for wild-type RUNX1 in leukaemogenesis have also been documented. For example, in AML patients with *FLT3*-ITD signalling mutations, constitutive MAP kinase activity leads to chronic activation of AP-1 proteins and their target genes, including the *RUNX1* locus. Genome-wide analyses of accessible chromatin and RUNX1 binding revealed that DHSs specifically activated by *FLT3*-ITD mutations were also occupied by RUNX1 and featured enrichment of AP-1 motifs, strongly suggesting AP-1 and RUNX1 functionally co-operate in response to *FLT3*-ITD signalling to promote leukaemogenesis[271]. Critically, this study also identified *FOXC1* as a downstream target of RUNX1 in *FLT3*-ITD AML patients, a gene which will be explored further in 1.5.6.

1.5.5 C/EBPα dysfunction also occurs frequently in the development of leukaemia

In addition to the indirect consequences of RUNX1 affecting C/EBPα function, the *CEBPA* gene is also mutated in ~10% of AMLs[290]. The most common *CEBPA* mutations with deleterious effects occur in the N-terminal region, generating frameshifts that abrogate the full-length 42 kDa (p42) isoform but still permit translation of a shorter 30 kDa (p30) isoform initiated downstream[291].

Expression of the p30 isoform of C/EBPα alone is sufficient to form GMPs in mice, but p42 is important to restrict excessive proliferation of the myeloid progenitor compartment. This function of p42 is underlined in mouse knockout studies, where targeted deletion of the p42 isoform alone gave rise to GMPs with aberrant self-renewal and subsequently a fully penetrant AML[292]. In AML, the p30 isoform alone exerts a dominant negative function over wild-type C/EBPα, causing downregulation of normal myeloid target genes[292,293]. A recent study also found the p30 isoform antagonises wild-type C/EBPα to de-repress the oncogenic lncRNA *UCA1*, droving proliferation of AML cells by downregulating the cell cycle inhibitor *CDKN1B*[294].

1.5.6 An emerging role for FoxC1 in AML

FoxC1 belongs to the Forkhead family of TFs introduced in 1.3.6, and features the winged-helix DNA binding domain common to all family members[122]. *In vitro* studies have demonstrated that FoxC1 preferentially recognises the 5'-TGTTTAR-3' variant of the Forkhead consensus sequence[295].

In embryogenesis, expression of FoxC1 or its closely-related paralogue FoxC2 at the correct level is essential for normal development of the ocular, skeletal, renal and cardiovascular systems. These roles are demonstrated in *Foxc1* and *Foxc2*-null mice, which are embryonic lethal, or conditional knockout animals which die at birth or soon after with skeletal, cardiovascular, renal and eye defects, and hydrocephalus[296]. Consistent with observations in mice, humans with heterozygous *FOXC1* mutation or deletion have haploinsufficiency phenotypes including Axenfeld-Rieger syndrome, a heritable ocular disorder, and congenital hydrocephalus[297,298]. In adult cells, FoxC1 expression is predominantly restricted to mesenchymal lineages including the bone and eyes, but also hair follicle stem cells, where it promotes quiescence[299].

Interestingly, *Foxc1* expression is also necessary for correct formation of the endosteal HSC niche in mice, but is notably not significantly expressed in any normal haematopoietic cell type[300,301]. Inducible deletion of *FoxC1* in both embryos and adults led to a depletion of HSPCs, strongly suggesting that its expression is necessary for maintenance of the HSC niche. Critically, FoxC1 is not normally expressed in any haematopoietic lineage in both mice and humans.

FoxC1 is frequently overexpressed in many solid malignancies, including those of the breast, liver, pancreas and skin[302–305]. In all of these cases, higher *FOXC1* mRNA expression correlates with an unfavourable prognosis and is associated with the epithelial-to-mesenchymal transition[306]. Critically, aberrant expression of *FOXC1* was more recently identified to be a feature of AML development. In one study, accessible chromatin and transcriptomic profiling of highly-purified AML patient blast cells found that *FOXC1* was specifically upregulated in patients with *FLT3*-ITD signalling mutations[271]. In parallel analyses, the authors of this study also identified a subset of DHSs upregulated in a FLT3-ITD-dependent manner and associated with FLT3-ITD-specific upregulated genes that were specifically enriched for occupied FOX:E-box (FOX:E) composite sites, increasing confidence that FoxC1 may play a direct role in the activation of leukaemia-specific genes.

At the same time, work from the Somervaille laboratory found that inappropriate *FOXC1* expression occurred more widely in AML patients with other mutations[300]. Based on analyses of large AML patient cohort RNA expression data, *FOXC1* upregulation was associated not just with *FLT3*-ITD mutations, but also *NPM1* signalling mutations and t(6;9) chromosomal translocations. Significantly, *FOXC1* expression in human AML strongly correlated with high HOX gene expression, in particular that of *HOXA9*.

1.5.7 FoxC1 enhances the clonogenic potential of human AML cell lines

The functional significance of *FOXC1* overexpression in AML has been validated by experiments in cell lines. shRNA-mediated depletion of *FOXC1* expression in various AML cell lines caused a reduction in clonogenic potential resulting from G1 cell cycle arrest and induction of morphological differentiation. Furthermore, forced differentiation of HL-60 cells into either monocytes or granulocytes was accompanied by a dramatic reduction in FoxC1 protein[300].

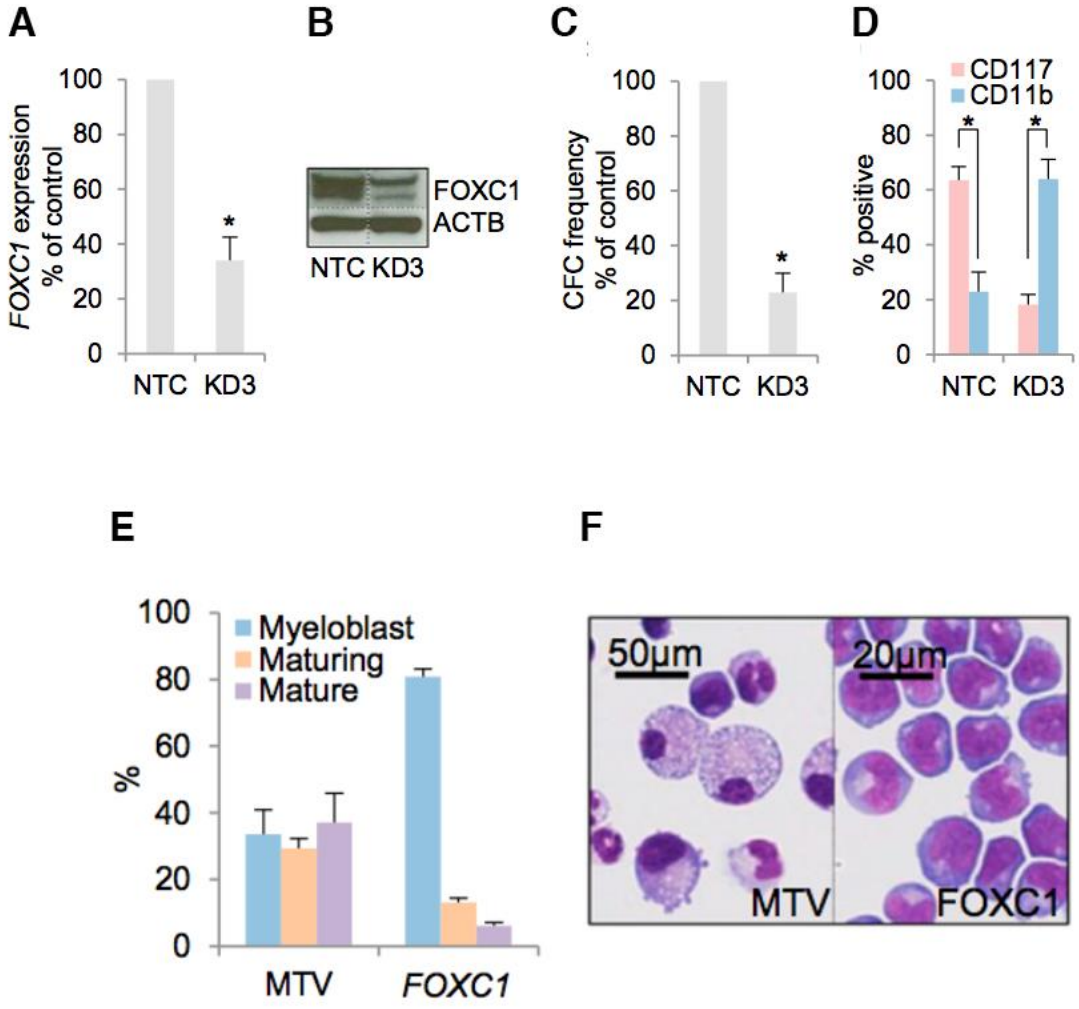


Figure 1:20 **Phenotypic interrogation of *FOXC1* in the context of AML**.

A-B: shRNA mediated knockdown of *FOXC1* in THP-1 AML cells depletes both mRNA and protein expression. In the Western blot analyses of protein expression shown in **B**, β-actin (ACTB) is used as a loading control.

C: Depletion of *FOXC1* was associated with a significant reduction in colony forming frequency of THP-1 cells.

D: Loss of the Leukaemia Stem Cell (LSC) marker CD117 surface expression following depletion of *FOXC1* was associated with a parallel upregulation of the myeloid marker CD11b.

E-F: Forced expression of *FOXC1* in mouse BM HSPCs induced a [transient] myeloid differentiation block.

From Somerville *et al.* (2015)

1.5.8 FoxC1 collaborates with HoxA9 to drive leukaemogenesis in mice

In mice, forced expression of *FOXC1* in normal BM HSPCs induced a transient myeloid differentiation block and enhanced proliferation, but was not sufficient to induce leukaemia[300]. Furthermore, it has previously been reported that *HoxA9* overexpression in mouse HSPCs alone generates a myeloproliferative disorder, but is insufficient for full leukaemic transformation[307]. However, co-expression of *FOXC1* with *HoxA9* in BM HSPCs was sufficient to generate aggressive serially-transplantable leukaemias in sub-lethally irradiated recipient mice (Figure 1:21).

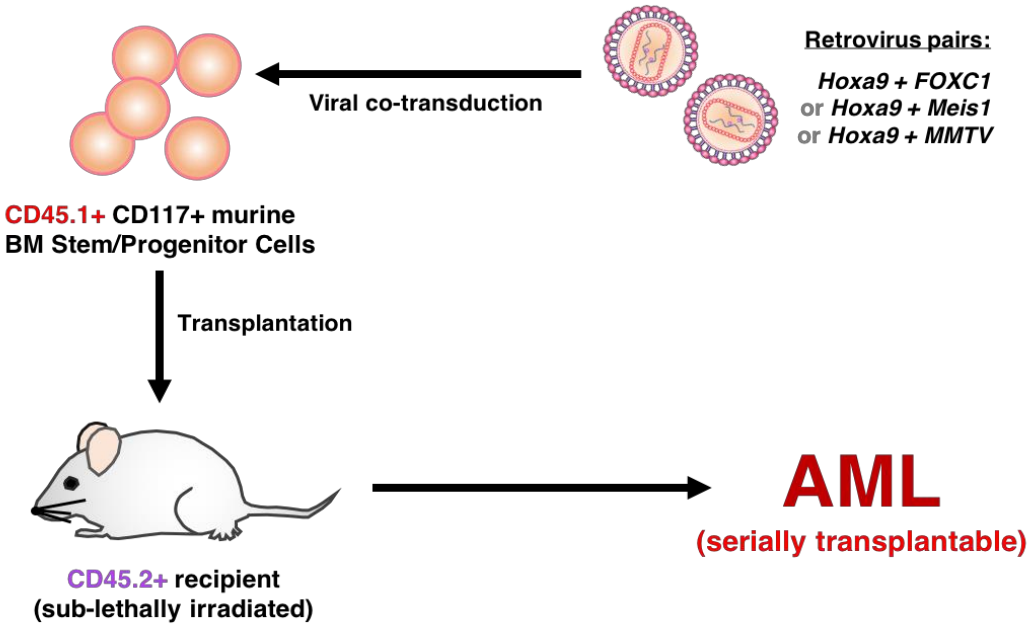


Figure 1:21 **Overview of the Somervaille Hoxa9+FoxC1-dependent AML mouse model.**

This schematic illustrates the key aspects of the syngeneic transplantation model used to gain extensive phenotypic insights into the consequences of forced *Hoxa9* and *FOXC1* expression in mouse BM HSPCs presented in Somerville *et al.* (2015).

Leukaemias in these mice were characterised by a monocyte differentiation block and extensive leukaemic blast cell infiltration of the bone marrow, liver and spleens. Furthermore, *HoxA9/FOXC1* AML cells were determined to be more enriched for actively cycling cells in the S/G2/M phases of mitosis compared to leukaemic cells from the well-defined *Hoxa9*/*Meis1* mouse model.

The authors also determined that *Klf4,* a TF associated with terminal myeloid differentiation, was downregulated in *Hoxa9/FOXC1* mice, and showed in human AML cells that knockdown of *FOXC1* led to restoration of *KLF4* expression. Conversely, forced expression of *KLF4* in *Hoxa9/FOXC1* AML cells significantly decreased their colony-forming capacity, suggesting downregulation of *Klf4* to be an important event in leukaemic transformation. However, despite these analyses revealing *Klf4* and general repression of monocyte/macrophage genes to be a consequence of *Hoxa9* and *FOXC1* over-expression, the direct targets of *FOXC1* in this context remained unknown.

1.6 Aims and Objectives

Previous studies of the Hoxa9+FoxC1-dependent mouse AML model clearly demonstrated a robust phenotypic contribution made by FoxC1 during leukaemogenesis. However, these studies did not extend to a full characterisation of the gene network activated in response to Hoxa9+FoxC1. Therefore, a primary aim of this work was to perform an integrative analysis of accessible chromatin, genomic occupancy and gene expression in Hoxa9+FoxC1 cells in order to gain an understanding of potential pathways and target genes that are modulated in a FoxC1-dependent manner. These aims are explored in the first results chapter.

A second aim of this work was to identify suitable cell models in which to study the consequences of FoxC1 expression in human AMLs. To address this, we performed a detailed analysis of FUJIOKA AML cells, which were found to express comparatively high levels of *FOXC1*. This is the focus of the second results chapter.

Finally, although previous studies in this laboratory suggested that *FOXC1* may contribute to the activation of FOX:E composite sites within DHSs in primary FLT3-ITD AMLs, these putative *cis*-regulatory elements had not been functionally interrogated. Thus, the principal aim of the work presented in the third results chapter was to use cell line models to further explore the potential significance of FOX:E composite DHSs in the activation of leukaemia-specific genes.

Chapter 2: Materials and Methods

Table 2:1 **Specialist Lab Equipment**

Product	Manufacturer	Software (if applicable)
ABI 7500 real-time PCR System	Applied Biosystems	StepOne™ Plus
Bioruptor™	Diagenode	
CyAn™ ADP flow cytometer	Beckman Coulter	Summit v4.3
Dounce homogeniser (15 ml)	Sigma Aldrich	
DynaMag™-2 magnetic rack	Thermo Fisher Scientific	
FACSAria™ II	BD Biosciences	
GD 2000 Gel Drier	Hoefer	
Gel Doc XR+	Bio-Rad	Image Lab™
Gene Pulser Xcell	Bio-Rad	
GloMax® Multi+ Microplate Multimode Reader	Promega	
HiSeq 2500	Illumina	
MiniMACS™ column	Miltenyi Biotec	
Multitron Standard Programmable Microbial Incubater	Infors HT	
NanoDrop™ 2000	Thermo Fisher Scientific	NanoDrop™ 2000
NextSeq 500	Illumina	
Optima™ MAX-XP Benchtop Ultracentrifuge	Beckman Coulter	
Orbital Shaking Incubator SSL1	Stuart Lab	
Personal Molecular Imager	Bio-Rad	Quantity One®
Prime Gradient Thermocycler	Techne	
Storage Phosphor Screen	Kodak	
T3000 96-well Thermocycler	Biometra	
ThermoMixer® C	Eppendorf	
TLN-100 Near-Vertical Ultracentrifuge Rotor	Beckman Coulter	
Trans-blot Turbo Transfer System	Bio-Rad	

Table 2:2 **Reagents and Suppliers**

Product	Manufacturer
2100 High Sensitivity DNA Kit	Agilent
Acetic Acid	VWR
Acrylamide/Bisacrylamide 37.5:1, 40%	Bio-Rad
Adenoside Triphosphate (ATP)	New England Biolabs
Agar	Thermo Fisher Scientific
Agarose	Thermo Fisher Scientific
Amersham ECL Western Blotting Detection Reagent Kit	GE Healthcare
Ammonium Persulphate	Sigma Aldrich
Ampicillin	Thermo Fisher Scientific
AMPure® XP beads	Beckman Coulter
Aprotinin	Sigma Aldrich
BamHI	New England Biolabs
BbsI	New England Biolabs
BglII	New England Biolabs
Bradford Reagent, Ready to Use	Bio-Rad
Bromophenol blue	Sigma Aldrich
BSA (Bovine Serum Albumin)	New England Biolabs
Buffer 2.1	New England Biolabs
Caesium Chloride (CsCl)	Sigma Aldrich
Calcium Chloride	Sigma Aldrich
Calcium ionophore, A23187 (CaI)	Sigma Aldrich
Calf Intestinal Phosphatase	New England Biolabs
Carbenicillin	Selleck Chemicals
CD117 MicroBead kit	Miltenyi Biotec
CD34 MicroBead kit	Miltenyi Biotec
Centricon® Plus-70 (100kDA NMWL)	Millipore
Cetyl Trimethylammonium Bromide (CTAB)	Sigma Aldrich
Chloroform	VWR
Citric Acid	Sigma Aldrich
cOmplete™ Protease Inhibitor Cocktail	Roche
Cutsmart™Buffer	New England Biolabs
Deoxyadenoside triphosphate (dATP)	New England Biolabs
Deoxycytosine triphosphate (dCTP)	New England Biolabs
Deoxyguanosine triphosphate (dGTP)	New England Biolabs
Deoxynucleotide (dNTP) solution mix	New England Biolabs
Deoxythymidine triphosphate (dTTP)	New England Biolabs
Disodium hydrogen phosphate (Na_2HPO_4)	Sigma Aldrich
Disuccinimidyl glutarate (DSG)	Sigma Aldrich
Dithiothreitol (DTT)	Sigma Aldrich
DNase I	Worthington
Dual-Luciferase Repoter Assay System Kit	Promega
Dynabeads™-Protein G	Thermo Fisher Scientific
EcoRI	New England Biolabs
Ethanol	VWR

Ethidium Bromide	Thermo Fisher Scientific
Ethylene-bis(oxyethylenenitrilo)tetraacetic acid (EGTA)	Sigma Aldrich
Ethylenedinitrilotetraacetic acid (EDTA)	Sigma Aldrich
Foetal Calf Serum	Thermo Fisher Scientific
Formaldehyde, 16%	Thermo Fisher Scientific
Glutamine	Thermo Fisher Scientific
Glycerol	Sigma Aldrich
Glycine	Sigma Aldrich
HEPES	Thermo Fisher Scientific
HyClone™ Ultra-pure water	Thermo Fisher Scientific
Hydrochloric Acid	VWR
Hyper Prep Kit (Illumina)	KAPA Biosystems
Illumina Library Quantification Kit	KAPA Biosystems
Illustra Probequant G-50 Spin Columns	GE Healthcare
IMDM	Thermo Fisher Scientific
Isopropanol	VWR
JM109 (*E. coli*)	Promega
LB Broth Base	Sigma Aldrich
Leupeptin	Sigma Aldrich
Lithium Chloride	Sigma Aldrich
Lymphoprep™	STEMCELL Technologies
MACS® buffer	Miltenyi Biotec
Magnesium Chloride	Sigma Aldrich
Marvel Dried Skimmed Milk Powder	Premier Foods
MicroPlex Library Preparation™ Kit v2	Diagenode
MinElute® Gel Extraction Kit	Qiagen
MinElute® PCR Purification Kit	Qiagen
MiniMACS™column	Miltenyi Biotec
Mini-PROTEAN® TGX Stain-Free Precast Gel, 4-20%	Bio-Rad
n-Butanol	VWR
NP-40	Sigma Aldrich
NucleoBond®Xtra Midi EF	Macherey Nagel
NucleoSpin® RNA kit	Macherey Nagel
OptiMEM™	Thermo Fisher Scientific
Penicillin/Streptomycin, 1%	Thermo Fisher Scientific
Phenol	VWR
Phenylmethylsulfonyl fluoride (PMSF)	Sigma Aldrich
Phorbol 12-myristate 13-acetate (PMA)	Sigma Aldrich
Phosphate buffered saline, sterile (PBS)	Thermo Fisher Scientific
PhosStop™complete phosphatase inhibitor	Roche
Phusion® High-fidelity PCR master mix	New England Biolabs
Poly(deoxyinosinic-deoxycytidyic) acid (polyIC)	Sigma Aldrich
Polynucleotide Kinase (PNK)	New England Biolabs
Polynucleotide Kinase Buffer	New England Biolabs
Potassium Chloride	Sigma Aldrich
Proteinase K	Sigma Aldrich

Purple Gel Loading Buffer, 6X	New England Biolabs
QIAprep® spin miniprep kit	Qiagen
QIAquick® gel extraction kit	Qiagen
Rainbow Molecular Weight Marker (High-Range)	GE Healthcare
RNase A	Worthington
RPMI+GlutaMAX™	Thermo Fisher Scientific
SOC medium	Thermo Fisher Scientific
Sodium Acetate (NaAc)	Sigma Aldrich
Sodium bicarbonate ($NaHCO_3$)	Sigma Aldrich
Sodium Chloride (NaCl)	Sigma Aldrich
Sodium deoxycholate	Sigma Aldrich
Sodium dodecyl sulfate (SDS)	Sigma Aldrich
Sodium Fluoride (NaF)	Sigma Aldrich
Sodium orthovanadate (Na_3VO_4)	Sigma Aldrich
Sodium β-glycerophosphate (Na_2 β-glycerophosphate)	Sigma Aldrich
Spectra/Por MWCO 3500 dialysis tubing	Spectrum Labs
Sucrose	Sigma Aldrich
SuperScript II Reverse Transcriptase Kit	Thermo Fisher Scientific
SYBR® Green PCR master mix	Thermo Fisher Scientific
T4 DNA ligase	New England Biolabs
T4 DNA ligase buffer	New England Biolabs
T4 DNA Polymerase	New England Biolabs
TE, 1X	Thermo Fisher Scientific
Terrific Broth Base	Sigma Aldrich
Tetramethylethylenediamine (TEMED)	Sigma Aldrich
Trans-Blot® Turbo™ Mini Nitrocellulose Transfer Packs	Bio-Rad
TransIT®-293	Mirius
Tris base	Sigma Aldrich
Triton™ X-100	Sigma Aldrich
TRIzol®	Thermo Fisher Scientific
TruSeq RNA sample preparation kit, v2	Illumina
Trypan Blue Solution, 0.4%	Thermo Fisher Scientific
TWEEN® 20	Sigma Aldrich
Whatman® 3MM blotting paper	Sigma Aldrich
XhoI	New England Biolabs
α-^{32}P-dCTP	Perkin Elmer

Table 2:3 **Buffer compositions**

Buffer	Composition
SDS Running Buffer (1X)	25 mM Tris-HCl pH 8.0 192 mM Glycine 0.1 % (v/v) SDS
TAE (1X)	40 mM Tris-Acetate 1 mM EDTA
DNase I Nuclei Digestion Buffer (2X)	120 mM KCl 30 mM NaCl 10 mM $MgCl_2$ 20 mM Tris-HCl pH 7.4
Western Blot Sample Loading Buffer (2X)	0.1 M Tris-HCl pH 8.0 0.2 M DTT 4 % (v/v) SDS 0.3 µM Bromophenol blue 20 % (v/v) Glycerol
EMSA Binding Buffer (4X)	20 % (v/v) Glycerol 60 mM HEPES 20 mM KCl 45 mM NaCl 20 µg/ml Aprotinin 20 µg/ml Leupeptin
Polyacrylamide Gel (4%) for EMSA	25 mM Tris base 0.5 mM EDTA 25 mM Boric Acid 4 % (v/v) Acrylamide/bisacrylamide (37.5:1) 0.06% (v/v) Ammonium Persulphate 0.3% (v/v) TEMED
TBE Buffer (5X)	250 mM Tris base 5 mM EDTA 250 mM Boric Acid
EMSA phosphatase inhibitor cocktail (8X)	8 mM DTT 0.8 mM PMSF 80 mM NaF 80 mM Na_2 β-glycerophosphate 8 mM Na_3VO_4
Agarose gel for electrophoresis	x % (w/v) Agarose 0.5 µg/µl Ethidium Bromide 40 mM Tris-Acetate 1 mM EDTA
ChIP Buffer A	10 mM HEPES pH 8.0 10 mM EDTA pH 8.0 0.5 mM EGTA pH 8.0 0.25 % (v/v) Triton™X-100 0.1 mM PMSF 1:100 (v/v) cOmplete™ Protease Inhibitor Cocktail
ChIP Buffer B	10 mM HEPES pH 8.0 0.2 M NaCl

	1 mM EDTA pH 8.0 0.5 mM EGTA pH 8.0 0.01 % Triton™X-100 0.1 mM PMSF 1:100 (v/v) cOmplete™ Protease Inhibitor Cocktail
ChIP Elution Buffer	100 mM $NaHCO_3$ 1 % (v/v) SDS
ChIP IP Buffer 1	25 mM Tris-HCl pH 8.0 150 mM NaCl 2 mM EDTA pH 8.0 0.25 % (v/v) SDS 0.1 mM PMSF 1:100 (v/v) cOmplete™ Protease Inhibitor Cocktail
ChIP IP Buffer 2	25 mM Tris-HCl pH 8.0 150 mM NaCl 2 mM EDTA pH 8.0 7.5 % (v/v) Glycerol 0.1 mM PMSF 1:100 (v/v) cOmplete™ Protease Inhibitor Cocktail
ChIP LiCl Buffer	10 mM Tris-HCl pH 8.0 250 mM LiCl 1 mM EDTA pH 8.0 0.5 % (v/v) NP-40 0.5 % (v/v) Na deoxycholate
ChIP TE/NaCl Buffer	10 mM Tris-HCl pH 8.0 50 mM NaCl 1 mM EDTA pH 8.0
ChIP Wash Buffer 1	20 mM Tris-HCl pH 8.0 150 mM NaCl 2 mM EDTA pH 8.0 1 % (v/v) Triton™X-100 0.1 % (v/v) SDS
ChIP Wash Buffer 2	20 mM Tris-HCl pH 8.0 0.5 M NaCl 2 mM EDTA pH 8.0 1 % Triton X-100 0.1 % (v/v) SDS
Citrate Phosphate Buffer	50 mM Na_2HPO_4 pH to 5.6 with Citric Acid
DNase I cell lysis buffer	0.3 M NaAc 10 mM EDTA pH 7.4 1 % (v/v) SDS
DNase I resuspension buffer	60 mM KCl 15 mM NaCl 5 mM $MgCl_2$ 10 mM Tris-HCl pH 7.4 0.3 M Sucrose

EMSA buffer A	10 mM HEPES 10 mM KCl 1.5 mM $MgCl_2$ 0.5 mM DTT 5 µg/ml Aprotinin 5 µg/ml Leupeptin 0.5 mM PMSF 1X PhosStop
EMSA buffer C	20 mM HEPES 0.42 M NaCl 1.5 mM $MgCl_2$ 0.2 mM EDTA 0.5 mM DTT 5 µg/ml Aprotinin 5 µg/ml Leupeptin 0.5 mM PMSF 10 mM NaF 10 mM Na_2 β-glycerophosphate 1 mM Na_3VO_4
EMSA buffer D	20 mM HEPES pH 7.9 100 mM KCl 0.5 mM DTT 20% (v/v) Glycerol 0.5 mM PMSF 1 µg/ml Leupeptin 10 mM NaF 10 mM Na_2 β-glycerophosphate 1 mM Na_3VO_4
Maxi-prep Alkaline Lysis Buffer	1 % (v/v) SDS 200 mM NaOH
Maxi-prep CsCl/TE Buffer	50% (w/v) CsCl 0.5 X TE
Maxi-prep Neutralisation Buffer	3 M KAc 2 M Acetic Acid
Maxi-prep Resuspension Buffer	50 mM Tris-HCl pH 8.0 10 mM EDTA
Western Blot Blocking Buffer	5% (w/v) Marvel Milk Powder 10 mM Tris-HCl pH 7.5 15 mM NaCl
Western Blot Washing Buffer	20 mM Tris-HCl pH 8.0 150 mM NaCl 0.1% TWEEN® 20
LB/amp	2 % (w/v) LB Broth base 100 µg/µl Ampicillin
TAE/EtBr	40 mM Tris-Acetate 1 mM EDTA 0.5 µg/ml Ethidium Bromide
LB agar/amp	2 % (w/v) LB Broth base 1.5 % (w/v) Agar 100 µg/µl Ampicillin

TB/carb	4.76% (w/v) Terrific Broth base 0.8 % (v/v) Glycerol 100 µg/µl Carbenicillin

Table 2:4 **Antibodies**

Antibody	Manufacturer	Catalogue	Experimental Conditions
Flow Cytometry			
Mouse anti-mouse CD45.1 (e450-conjugated)	Thermo Fisher Scientific	47-0453-82	1:100
Mouse anti-mouse CD45.2 (e780-conjugated)	Thermo Fisher Scientific	47-0454-82	1:100
Mouse anti-mouse CD11b (PE-Cy7-conjugated)	Thermo Fisher Scientific	25-0112-82	1:100
Mouse anti-mouse Gr1 (FITC-conjugated)	Thermo Fisher Scientific	11-5931-82	1:100
Mouse anti-mouse B220 (APC-conjugated)	Thermo Fisher Scientific	17-0452-82	1:100
Mouse anti-mouse CD43 (PE-conjugated)	Thermo Fisher Scientific	12-0431-82	1:100
Mouse anti-mouse CD4 (PE-Cy5-conjugated)	Thermo Fisher Scientific	15-0042-82	1:100
Mouse anti-mouse CD8 (APC-conjugated	Thermo Fisher Scientific	17-0081-82	1:100
Mouse anti-human CD34 (FITC-conjugated)	BioLegend	343504	1:100
Mouse anti-human CD117 (PE-conjugated)	BioLegend	313204	1:100
ChIP-seq			
Goat anti-human FoxC1	Abcam	ab-5079	5 µg per IP (overnight)
Goat anti-human C/EBPα	Santa Cruz	sc-61-X	1 µg per IP (2 h)
Rabbit anti-human RUNX1	Abcam	ab23980	2 µg per IP (overnight)
EMSA Supershift Assays			
Goat anti-human FoxC1	Abcam	ab-5079	1 µg per assay
Goat anti-human FoxC1	Santa Cruz	sc-2139-X	1 µg per assay
Rabbit anti-human FoxO3a	Cell Signalling Technology	75D8	1 µg per assay
Mouse anti-human FoxP1	Santa Cruz	sc-398811-X	1 µg per assay
Mouse anti-human IgG	Santa Cruz	sc-66931-X	1 µg per assay
Rabbit anti-human E2A	Santa Cruz	sc-763-X	1 µg per assay
Goat anti-human TCF4 (E2-2)	Santa Cruz	sc-8631-X	1 µg per assay
Western Blotting			

Rabbit anti-human GAPDH	Cell Signalling Technology	14C10	1:4000 per membrane (overnight, 4°C)
Rabbit anti-human FoxO3a	Cell Signalling Technology	75D8	1:500 per membrane (overnight, 4°C)
Mouse anti-Rabbit IgG (HRP-conjugated secondary)	Rockland	18-8816-31	1:2000 (1 h, RT)

Table 2:5 **Expression vector backbones**

Vector Backbone	**Source**
Luciferase Reporter Gene Assays	
pXPG-TK229	Bert *et al.* (2000) *Plasmid* **44**(2)
pRL-TK750	Promega
CRISPR/Cas9 Genome Editing	
pX458-sSpCas9(BB)-2A-GFP	Ran *et al.* (2013) *Nature Protocols* **8**(11)
shRNA-mediated knockdown	
LeGO-iG	Weber *et al.* (2008) *Molecular Therapy* **16**(4)
Tat	Sommer *et al.* (2009) *Stem Cells* **27**(3)
Rev	
Gag/Pol	
VSV-G	

Table 2:6 **Oligonucleotides**

Name	**Forward Sequence (5' - 3')**	**Reverse Sequence (5' - 3')**
Gene Expression – qPCR primers		
hFOXC1	CCCTCTCTTGCCTTCTTCCT	CGTCAGGTTTTGGGAACACT
hGAPDH	CCTGGCCAAGGTCATCCAT	AGGGGCCATCCACAGTCTT
DNase I – qPCR validation primers		
Human		
***TBP* promoter**	CTGGCGGAAGTGACATTATCAA	GCCAGCGGAAGCGAAGTTA
***ACTB* gene body**	GCAATGATCTGAGGAGGGAAGGG	GTGTCTTTCCTGCCTGAGCTGAC
Chr18	ACTCCCCTTTCATGCTTCTG	AGGTCCCAGGACATATCCATT
***TBP* promoter (short amplicon)**	CTGGCGGAAGTGACATTATCAA	CCCGACCTCACTGAACCC
***ACTB* gene body (short amplicon)**	GCAATGATCTGAGGAGGGAAGGG	AGCTGTCACATCCAGGGTCCTCA

Chr18 (short amplicon)	AGGTCCCAGGACATATCCATT	GTTCAAATTGTGTTTTGTGGTTA
Mouse		
***Tbp* promoter**	TGCAGTCAAGAGCGCAACTG	CACCGCTACCGGACTCGAT
***Actb* gene body**	CAAGACAAGATGGTGAATGGTGAG	GTAAGCTAAGCATCCTTAGCTTGG
Chr1	TGCTCCACAGTGTCCATGTACA	AGCAATTTCATGGGTGAGAGAAG
***Tbp* promoter (short amplicon)**	GGCGGAAGTGACGGCATCAG	CACCGAACGCCTGACGCAC
***Actb* gene body (short amplicon)**	GGTGAGCTCTCTGGGTGC	GGTTGTCAGAGCAACCTTCTAGG
Chr1 (short amplicon)	GTGTCCATGTACAATTGAGTTTGG	GGTATACCTGGTGGGATTTC
ChIP-seq – qPCR validation primers		
FoxC1 ChIP – putatively positive regions		
***DUSP6* -130 kb DHS**	GGATACATTCTCCGGGCCAG	GTTTGCACTGGGGCTTATCC
***GTF2A1* promoter**	TTAAATGGAGAGGCGGTGAC	CCTTTGGCTGCCCTGTTTAT
C/EBPα and RUNX1 ChIP – validated positive regions		
***SPI1* -14 kb enhancer**	AACAGGAAGCGCCCAGTCA	TGTGCGGTGCCTGTGGTAAT
***IGFBP7* promoter**	GTCAAGCACTAAAAGGACAAACCG	TGAATGCCACTGGGAGACAAAG
Negative Control regions for all ChIP assays		
***IVL* promoter**	GCCGTGCTTTGGAGTTCTTA	CCTCTGCTGCTGCCACTT
hChr18	ACTCCCCTTTCATGCTTCTG	AGGTCCCAGGACATATCCATT
EMSA probes		
SCARA3	GATCCCAGAGGCTCTGTTTTGACAGGTGTGTCC	GATCGGACACACCTGTCAAAACAGAGCCTCTGG
SCARA3dE	GATCCCAGAGGCTCTGTTTTGACACTAATGTCC	GATCGGACATTAGTGTCAAAACAGAGCCTCTGG
SCARA3dF	GATCCCAGAGGCTCAACACTTACAGGTGTGTCC	GATCGGACACACCTGTAAGTGTTGAGCCTCTGG
SCARA3dEdF	GATCCCAGAGGCTCTGCCCTGACACTAATGTCC	GATCGGACATTAGTGTCAGGGCAGAGCCTCTGG
FAM92A1	GATCTAAAAGTTTGTGTTTACACAGGTGCAGCCA	GATCTGGCTGCACCTGTGTAAACACAAACTTTTA
FAM92A1dE	GATCTAAAAGTTTGTGTTTACAGAGGCACAGCCA	GATCTGGCTGTGCCTCTGTAAACACAAACTTTTA
FAM92A1dF	GATCTAAAAGTTTGTACATACACAGGTGCAGCCA	GATCTGGCTGCACCTGTGTATGTACAAACTTTTA
ALB	TACCAGGGAATGTTTGTTCTTAAATAC	GTATTTAAGAACAAACATTCCCTGGTA
eNOS	TAGGAGGGGGTGTTTACTGAAACTAGG	CCTAGTTTCAGTAAACACCCCCTCCTA
VANGL1	GATCTGAAGCCCTGTGTTTATCCAGCTGCCTC	GATCGAGGCAGCTGGATAAACACAGGGCTTCA
CLPS	AAAGCTTCTCTGTTTGCATACCGAAGG	CCTTCGGTATGCAAACAGAGAAGCTTT

Luciferase Reporter Gene Assays – DHS fragment cloning primers		
Fox:Ex3	CAAGGATCCTAAAGTTTGTGTTTACACAGGTGCAGCCATTAAAGTTTGTGTTTACACAGGTGCAGCCATTAAAGTTTGTGTTTACACAGGTGCAGCCATCTCGAGAAC	GTTCTCGAGATGGCTGCACCTGTGTAAACACAAACTTTAATGGCTGCACCTGTGTAAACACAAACTTTAATGGCTGCACCTGTGTAAACACAAACTTTAGGATCCTTG
FOXC1	GTGCGAGATCTGGAAGGCATGTTTCT	GGTGGCCAGATCTTTTGCCATCCCA
FAM92A1	TAAGAGATCTGATTGACAACTGACCAAAACCT	CTTAAGATCTGCTTGCACACATATTAGACGGT
VANGL1	TAAGAGATCTTGTCCAAGCCAGTGATTACAT	CTTAAGATCTAACACCTCCACTCCACCCA
SCARA3	TAAGAGATCTTCCTTCTAAGAAGAGCAGGGG	CTTAAGATCTCAGAGTCTTGCACTTGGAAGTG
IGFBP2	TAAGGGATCCTCTTCCTCCTACAAAGAACGTT	CTTACTCGAGAGCAGAGAGTCATGACATGGA
RASGRP3	TAAGAGATCTGACCTCACCACCTCTCCAA	CTTAAGATCTAGATTTGTTGGGTAGAAATGCC
IL2RA(DHS 1)	TAAGGGATCCGGTACCTTTGTCTTCTGAGTGC	CTTACTCGAGCACCAGGGATGAAATTAGCCT
IL2RA(DHS 2)	TAAGGGATCCGCTGGAACTATAGGCGTGCG	CTTACTCGAGAGTCATCCCCAACTGGCAA
General Plasmid Sequencing Primers		
pXPG sequencing primer	TATCGCTACGTGACTGGGTC	
U6 promoter sequencing primer	GAGCTCACAACCCCTCACTC	
CRISPR deletion of *FOXC1* – sgRNA cloning and PCR screening		
sgRNA 5' guide 1	CACCGAGAGCTACTACCGCGCGG	AAACCCGCGCGGTAGTAGCTCTC
sgRNA 5' guide 2	CACCGCTGCTCGCCGCCGAGGTA	AAACTACCTCGGCGGCGAGCAGC
sgRNA 3' guide 1	CACCGGATTTAGTTCGGCTTTGA	AAACTCAAAGCCGAACTAAATCC
sgRNA 3' guide 2	CACCGCGATTTAGTTCGGCTTTG	AAACCAAAGCCGAACTAAATCGC
PCR Primer 1	GTTTGCGCCTGGAAGCT	
PCR Primer 2	ATGGTGATGAGCGCGATGTA	
PCR Primer 3	CAGCAGCAGAACTTCCACTC	
PCR Primer 4	AGGTCTCTGAAAAGCAAGAAGA	
FoxC1 shRNA knockdown		
sh FOX B	TGTCACAGAGGATCGGCTTGAATTCAAGAGATTCAAGCCGATCCTCTGTGACTTTTTTC	TCGAGAAAAAAGTCACAGAGGATCGGCTTGAATCTCTTGAATTCAAGCCGATCCTCTGTGACA
sh MM	TGCGCGATAGCGCTAATAATTTTTCAAGAGAAAATTATTAGCGCTATCGCGCTTTTTTC	TCGAGAAAAAAGCGCGATAGCGCTAATAATTTTCTCTTGAAAAATTATTAGCGCTATCGCGCA

Table 2:7 **Specialist software used for Bioinformatics**

Software	Version
BEDtools	2.25.0
Bowtie	1.0.0
Bowtie2	2.3.1
Cutadapt	1.9.1
DAVID	6.8
HiSat2	2.10
Homer	4.9
Integrated Genomics Viewer	2.4
Java TreeView	3.0
MACS	1.4.2
pyBedtools	0.7.10
R	3.23
R package: ChIPpeakAnno	3.11.6
R package: Limma	3.28.14
R package: pheatmap	1.0.8
R package: VennDiagram	1.6.17
SAMtools	1.5
StringTie	1.33
Trimmomatic	0.36
Wellington	0.2.4

2.1 Cell line culture

2.1.1 General cell culture

Leukaemia cell lines were maintained in RPMI/GlutamaxTM supplemented with 10% heat-inactivated foetal calf serum and 1% penicillin/streptomycin in a humidified incubator at 37°C and 5% CO_2. All cell lines were maintained at densities recommended by the cell bank from which the cells originated (Table 2:8).

Table 2:8 **Cell lines used in this study**

Cell Line	Cell Bank/Accession Number
Jurkat	DSMZ/ACC-282
THP-1	ATCC/TIB-202
FUJIOKA	JCRB/0091
MV4-11	ATCC/CRL-9591
MOLM-14	DSMZ/ACC-777
PL-21	DSMZ/ACC-536
KG1a	ATCC/CCL-246.1
K562	ATCC/CCL-243
NB-4	DSMZ/ACC-207

2.1.2 Cell stimulation

A 100 X stock of calcium ionophore A23187 (CaI) and phorbol 12-myristate 13-acetate (PMA) was prepared in cell media immediately prior to stimulation. Cells were stimulated by addition of a 1/100 volume of this stock, to give a final concentration of 2 μM CaI and 20 ng/ml PMA.

Depending on the downstream analyses required, cells were stimulated for 4 hours (for nuclear extracts to use in EMSAs) or 8 hours (for luciferase reporter gene assays).

2.1.3 Preparation of conditioned medium for single cell CRISPR clone culture

FUJIOKA cells in exponential growth (i.e. with a density of 3.0-5.0x10^5/ml) were harvested from culture by centrifugation (400 xg, RT, 5 min). The conditioned medium was aspirated and passed through a 0.22 µm pore filter to remove residual cells.

Filtered conditioned medium was with fresh RPMI+Glutamax™ supplemented with 20% heat-inactivated foetal calf serum and 1% penicillin/streptomycin to give a final concentration of 40% conditioned medium for use in culture.

2.2 Processing of primary AML patient samples

Use of human tissue in this study was in compliance with the ethical and legal framework of the United Kingdom's Human Tissue Act of 2004.

Peripheral blood samples from primary AML patients were obtained from the University Hospitals Birmingham NHS Foundation Trust (source) and diluted to a maximum density of 1x10^7 cells/ml with sterile PBS. Blood samples were carefully overlaid onto an equal volume of Lymphoprep™ in 50 ml Falcon tubes prior to centrifugation (800 xg, RT, 20 min, brake off). The upper plasma layer was carefully discarded, then the plasma:Lymphoprep™ interface layer, containing peripheral blood mononuclear cells (PBMCs), was transferred to a new Falcon tube and diluted with 50 ml sterile PBS. Cells were recovered by centrifugation (500 xg, RT, 10 min, brake on).

AML blast cell purity was determined by flow cytometry analysis (2.4). If the surface expression of CD34 or CD117 was below 92%, cells were further purified using either a CD34 or CD117 MicroBead kit and a MiniMACS™ column, according to the manufacturer's instructions. Purified PBMCs were aliquoted for RNA-seq, DNase-seq and ChIP-seq analyses (2.5, 2.6 and 2.7).

2.3 Processing of murine AML samples

Cryopreserved spleen cells from murine AML samples generated by methods described in Somerville *et al.* (2015) were obtained from Tim Somervaille (Cancer Research UK Manchester Institute). Mouse experiments were approved by the Cancer Research UK Manchester Institute's Animal Ethics Committee and performed under a United Kingdom Home Office project, in adherence to the Animal Scientific Procedures Act of 1986.

Samples were thawed in 30 ml of IMDM with 10% heat-inactivated foetal calf serum, and a 5 µl aliquot was removed and mixed 1:1 with 0.4% Trypan blue solution to determine cell density and viability. Cell suspensions were aliquoted for subsequent RNA-seq and DNase-seq analyses (2.5 and 2.6).

2.4 Flow cytometry

$1x10^5$ cells per analysis were collected by centrifugation (400 xg, RT, 5 min). Pellets were washed twice in 1 ml of MACS buffer then re-suspended in 200 µl of supernatant. 2 µl of antibody was added and cell suspensions incubated in the dark at 4°C for 15 min. Cell suspensions were washed with a further 1 ml of MACS buffer prior to analysis with a CyAn™ ADP flow cytometer.

2.5 Analysis of gene expression

2.5.1 RNA isolation

$5x10^6$ cells were centrifuged (300 xg, 4°C, 5 min) then lysed by addition of 1 ml of TRIzol™ reagent. Samples were incubated at RT for 5 min prior to addition of 200 µl of chloroform and a further incubation at RT for 3 min. Samples were centrifuged (12000 xg, 4°C, 5 min) and the upper aqueous phase containing RNA transferred to a new tube. RNA was then purified using a NucleoSpin® RNA kit, with the optional on-column DNase digestion step, according to the manufacturer's instructions. RNA was quantified by 260 nm absorbance using a NanoDrop™ 2000.

2.5.2 cDNA synthesis

1 µg of total RNA was used in cDNA synthesis reactions with a SuperScript II Reverse Transcriptase kit and 0.5 µg of included Oligo(dT)$_{12-18}$ primers, according to the manufacturer's instructions.

2.5.3 Quantitative PCR

cDNA was quantified by real-time PCR in 10 µl reactions, using 200 nM of forward and reverse transcript-specific primers (Table 2:6) and SYBR® Green PCR master mix, according to the manufacturer's instructions. Analysis was performed on an ABI 7500 real-time PCR system using StepOne^{TM} Plus software.

Target gene expression was calculated relative to *GAPDH* expression, using transcript-specific standard curves generated using the following dilution series: 1:10, 1:50, 1:250, 1:1,250 and 1:6,250 of sample cDNA diluted with water. cDNA to be quantitated was diluted 1:10 or 1:50 times for low- or high-abundance transcripts, respectively.

2.5.4 RNA-seq library preparation

RNA-seq libraries were generated from 500 ng of total RNA using a TruSeq RNA sample preparation kit v2, following the manufacturer's instructions. Library synthesis efficiency was assessed using a 2100 High Sensitivity DNA Kit and library concentration determined using an Illumina Library Quantification Kit, both according to the manufacturer's instructions. Libraries were pooled with complementary indices at equimolar ratios for multiplexed paired-end sequencing on a NextSeq 500 high-throughput sequencer by the University of Birmingham Genomics Facility.

2.5.5 RNA-seq data processing

Raw paired-end reads were processed with Trimmomatic[308] to remove low-quality reads and adapter sequences. Reads were then mapped to the mouse genome assembly mm10 or human genome assembly hg19 using HiSat2[309], using default parameters.

RefSeq transcript abundances were calculated as Transcripts Per Million (TPM) using StringTie[309]. Only genes with a TPM > 1 in at least one sample were considered expressed and were retained for further analyses.

2.5.6 Correlation clustering of RNA-seq data

Hierarchical clustering of samples based on their gene expression profile was carried out using log-transformed TPM values (log2 TPM + 1). Pearson correlation values were calculated for each pair of samples using R. This correlation was then converted to a distance (1 – Pearson correlation), and clustered using average linkage hierarchical clustering. The result of this was plotted as a heatmap using the pheatmap package in R.

2.5.7 Differential gene expression analyses

Differential gene expression was conducted using the Limma package in R[310]. A gene was considered significantly differentially expressed if it had a fold-change > 2, and an adjusted *p* value < 0.05 between the two conditions tested. Adjusted p-values were calculated using the Benjamini-Hochberg method. KEGG pathway analyses were carried out using DAVID v6.8[311]. KEGG pathways with a Benjamini-Hochberg adjusted *p* value < 0.05 were considered significantly enriched.

2.6 DNase I hypersensitive site mapping

2.6.1 DNase I assay

2 - 3.5x10^7 cells were used per cell type tested, to allow a titration series of DNase I enzyme, and capture digestion points within the optimal range for efficient detection of both constitutive and inducible DHSs.

4.5x10^6 cells per DNase I digestion point were centrifuged (300 xg, RT, 5 min), washed in PBS and then re-suspended at 3x10^7 cells/ml in DNase I resuspension buffer. 150 µl aliquots of cells were equilibrated in 1.5 ml Eppendorf tubes to 22°C in a water bath for 3

min prior to nuclei digestion. Following this, an equal volume of 2X DNase I nuclei digestion buffer was added, with a final enzyme concentration generally ranging from 1 to 10 µg/ml, and samples were incubated at 22°C for exactly 3 min. Digestions were terminated by addition of 300 µl of DNase I cell lysis buffer, and samples were incubated overnight at 45°C.

2.6.2 Analysis of DNase I digestion points by agarose gel electrophoresis

100 µg/ml of RNase A was added to samples and incubated at 37°C for 1 h. 6 µl of each sample (1% of $4.5x10^6$ cells = ~225 ng DNA) were loaded with 1X Purple Gel Loading Buffer in an 18 µl volume onto a 0.8% agarose gel with 1X TAE buffer and 0.5 µg/ml ethidium bromide (TAE/EtBr). Electrophoresis was performed at 30 V for 16 h and DNA fragments visualised by UV trans-illumination using a Gel Doc XR+ and Image Lab software. A representative gel analyses is shown in Figure 2:1A.

2.6.3 DNA purification

The following morning, a further 100 µg/ml of RNase A was added to samples determined by gel electrophoresis analyses to possibly be close to the optimal DNase I digestion point, and incubated at 37°C for 30 min. An equal volume of phenol was added, and the mixture incubated with rotation at RT for 1 h. Samples were centrifuged (18,500 xg, RT, 5 min) and the upper aqueous phase transferred to new 1.5 ml Eppendorf tubes. An equal volume of 1:1 phenol:chloroform solution was added and incubated with rotation at RT for 30 min then centrifuged as previously. The aqueous phase was transferred to a new tube, an equal volume of chloroform added and incubated with rotation at RT for 30 min before centrifugation. The purified DNA was precipitated from the final aqueous phase by addition of two volumes of 100% ethanol and 50 mM NaCl, and stood on ice for 15 min.

Precipitated DNA was pelleted by centrifugation (18,500 xg, 4°C, 15 min), pellets washed in 500 µl 75% ethanol then recovered by centrifugation (18,500 xg, 4°C, 5 min). The supernatant was discarded and pellets were air-dried for 10-15 min at RT prior to re-suspension in 30 µl of 1X TE. Samples were stood overnight to allow complete dissolution, then quantitated by 260 nm absorbance using a NanoDrop™ 2000 spectrophotometer.

2.6.4 DNase-seq library preparation

Favourable digestion points were identified by qPCR analyses (2.5.3), using DNase digestion points as the template and genomic DNA for standard curves. The ratio of PCR signal was determined between a previously defined active region at the *TBP* promoter, which was expected to be DNase I hypersensitive, and an inactive region at either chromosome 1 (Chr1, for murine samples) or chromosome 18 (Chr18, for human samples), which was not expected to be hypersensitive. Optimally digested samples gave an active/inactive ratio of between ~0.4 to 0.7. A representative example of these qPCR analyses are shown in Figure 2:1A.

≥ 5 µg of DNA from the most favourable samples were loaded with 1 X Purple Gel Loading Buffer on a 1.5% agarose gel with TAE/EtBr, and separated by electrophoresis at 80 V for 2 h. DNA fragments of ~80 - 250 bp were excised using a clean scalpel then purified using a MinElute® Gel Extraction kit, according to the manufacturer's instructions.

Size-selected DNA fragments were validated again by qPCR, this time using shorter amplicon PCR primers for control active regions at *TBP* and *ACTB* promoters, and control inactive Chr1/Chr18 loci as appropriate. Samples showing the highest *TBP/ACTB*

and *TBP*/inactive enrichments were selected for genome-wide sequencing. Figure 2:1B presents representative examples of these steps in the protocol.

DNase-seq libraries were prepared using a MicroPlex Library Preparation kit v2, according to the manufacturer's instructions. Amplified libraries were loaded with 1X green gel loading buffer and separated by electrophoresis on a 1.5% agarose gel with TAE/EtBr at 80 V for 2 h. Library fragments of ~170 - 300 bp were purified using a MinElute® Gel Extraction kit according to the manufacturer's instructions, and eluted in a final volume of 12 µl. Representative examples of DNase-seq library size selection are presented in Figure 2:1C.

2.6.5 Quality control and pooling of DNase-seq libraries

To assess library quality for sequencing, libraries were re-validated using the short amplicon primers by qPCR. If the active/inactive enrichment was deemed to be satisfactory, the average library fragment size was determined using a 2100 High Sensitivity DNA Kit and library concentration quantified with an Illumina Library Quantification Kit, according to the manufacturer's instructions. Representative examples of these steps are shown in Figure 2:1C.

Libraries were pooled with complementary indices at equimolar ratios for multiplexed single-end sequencing on a NextSeq 500 high-throughput sequencer (Illumina). For ultra-high depth sequencing as required for digital DNase I footprinting analyses, libraries were re-sequenced using a HiSeq 2500 ultra-high-throughput sequencer (Illumina). Sequencing services were provided by the University of Birmingham Genomics Facility.

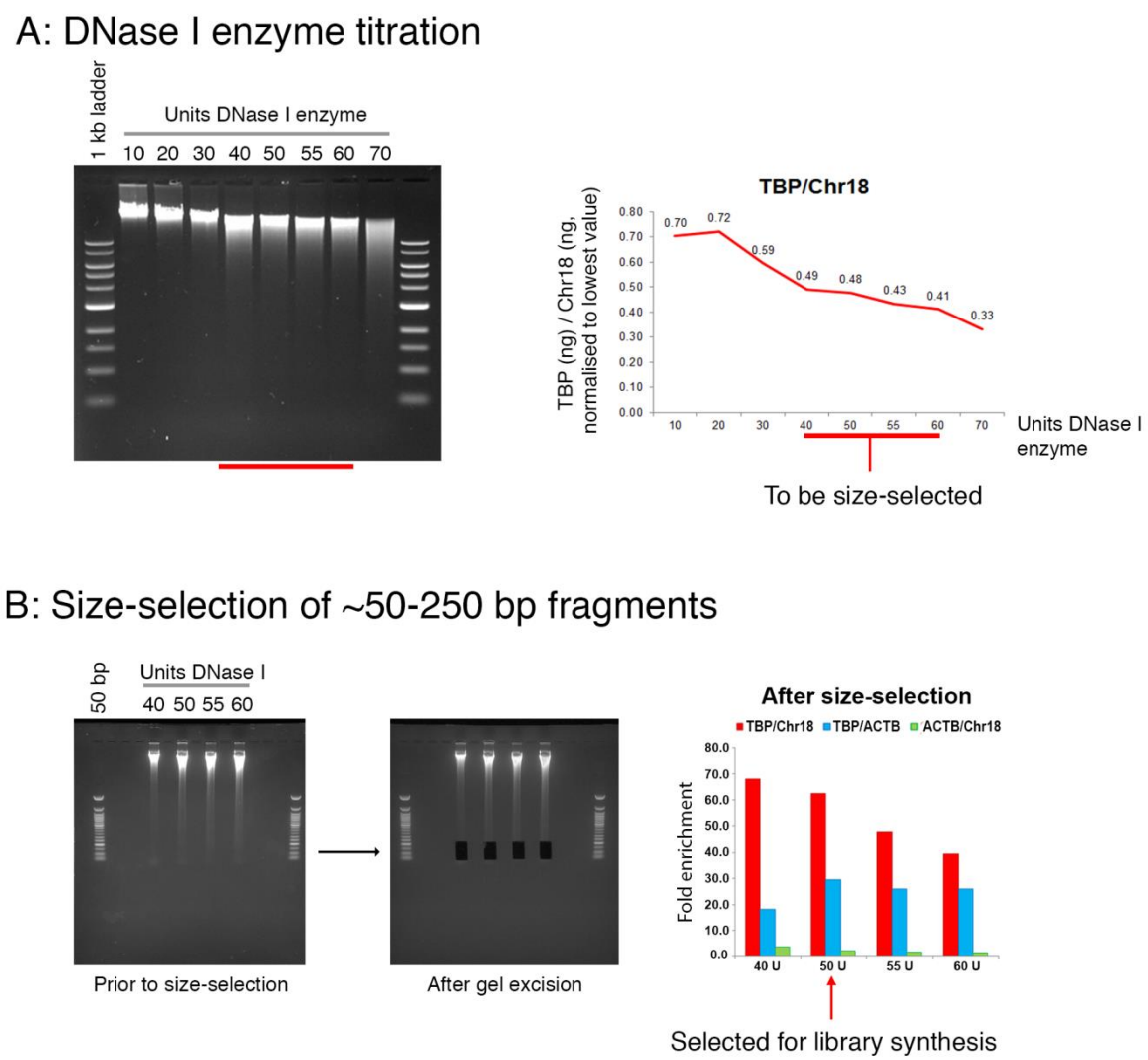


Figure 2:1 **Overview of DNase I digestion assay and sequencing library preparation steps as described in 2.6.1 to 2.6.5.**

2.6.6 DNase-seq data processing

Raw DNase-seq reads were processed with Trimmomatic[308] to remove low-quality reads and sequencing adapters. Reads were then mapped to the mouse genome assembly mm10 or human genome assembly hg19 using Bowtie2[312], using the parameter –very-sensitive-local. Regions of enrichment corresponding to open chromatin (peaks) were identified with MACS[313] using the options –g –keep-dup=all –w –S. Peaks were annotated to their closest gene using the annotatePeaks.pl function of Homer[314], and further annotated to a gene promoter if they were within 2 kb up or down-stream of the gene transcription start site (TSS) and as distal peaks if otherwise. Promoter and distal peaks were treated separately in further analyses.

Peak unions were constructed by merging peaks that had summits within 400 bp of each other using the merge function in BEDtools[315]. In these cases, a new summit position was calculated as the mid-point between the summits of the original peaks. This average summit position was used in further downstream analysis.

2.6.7 DNase-seq correlation analyses

Hierarchical clustering of DNase-seq data was carried out on the union of all peaks across all datasets. The number of mapped reads (tag density) for each peak was calculated for 400 bp windows centered on the peak summits using the annotatePeaks.pl function in Homer, and normalized by total tag count across all peaks. Tag counts were log-transformed (using log2 tag-count + 1) prior to calculation of Pearson correlation values for each pair of samples in R. These correlation values were then converted to a distance (1 – Pearson correlation value), and clustered using average linkage hierarchical clustering. The result of this clustering was plotted as a heatmap using the pheatmap package in R.

2.6.8 Identification of sample-specific DNase-seq peaks

To identify shared and specific sets of peaks, the tag density for each peak was calculated across 400 bp windows centred on the summits of the peak union, and normalized by total tag density in all peaks. A peak was considered specific to a condition if the normalized tag count was at least 4-fold higher in one condition relative to another, and as shared between conditions if otherwise.

2.6.9 DNase-seq motif enrichment analyse

A *de novo* motif search for TF binding motifs was conducted within the shared and specific populations of peaks using the findMotifsGenome.pl function in Homer, considering motif lengths of 8, 10 and 12 bp. Density plots were generated by ranking peaks according to their fold-difference, and then calculating the tag density across a 2kb window using annotatePeak.pl from Homer with the parameters –size 2000 –hist 10 –ghist followed by visualisation using Java TreeView[316]. Average profiles for the shared and specific peaks were plotted in R using the tag densities calculated by Homer, normalized by number of tags per million.

2.6.10 DNase-seq motif co-localisation analyses

Specific and shared populations of DHSs were identified from high read-depth DNase-seq experiments as described for low-depth samples. Genomic coordinates for TF binding motifs within these DHSs were retrieved using the annotatePeaks.pl function in Homer, and exported as BED files using the –mbed option. These motif searches were carried out using the motif probability weight matrices provided by the Homer database. Motif co-localization within the specific DHS populations was measured by counting the number of pairs of motifs within 50 bp of each other using the intersection_matrix function in pybedtools[317].

To assess the significance of co-localized motifs within the specific populations of DHSs, we carried out a bootstrapping analysis whereby a random set of DHSs were sampled with replacement from the union of all DHSs in Hoxa9+FoxC1 and Hoxa9+MMTV cells. This procedure was repeated 1000 times, and the number of co-localized motif pairs was counted in each of these bootstrap replicates. The mean and standard deviation of the counts across all replicates was then used to calculate a Z-score for each pair of motifs. Hierarchical clustering of the resulting Z-score matrix was performed using complete linkage clustering of the Euclidean distances, and plotted as a heatmap using the pheatmap package in R.

2.6.11 High-resolution digital DNase I footprinting analyses

Libraries corresponding to the highest-quality samples were re-sequenced on an Illumina HiSeq (University of Birmingham Genomics Facility) until a minimum number of 400,000,000 total reads per sample were obtained. Raw sequencing data from high-depth DNase-seq experiments were processed as described for low-depth samples.

Digital genomic footprints were identified with the Wellington algorithm (Piper et al. 2013) using the parameter --fdrlimit -10. Shared and specific sets of footprints were identified using the bedtools intersect function. A footprint was considered specific to an experimental condition if it did not overlap with a footprint in any other condition, and as shared if otherwise. A *de novo* motif search was carried out in specific footprints using Homer's findMotifsGenome.pl function with the parameter –size given. DHS cut tracks showing the strand imbalance in DNase I cut sites within footprinted regions were created using the dnase_wig_tracks.py function in Wellington, and visualized using the Integrated Genomics Viewer[318].

2.7 Chromatin Immunoprecipitation (ChIP)

2.7.1 Chromatin preparation

$3x10^7$ cells were centrifuged (400 xg, RT, 5 min) and pellets washed thrice with 50 ml PBS. The washed pellet was re-suspended to $1x10^7$ cells/ml with PBS and cross-linked with 0.83 mg/ml DSG on a rotating wheel at RT for 45 min. Single-crosslinked cells were washed three times, each with 50 ml PBS, then re-suspended in 15 ml PBS prior to addition of 1% formaldehyde (~0.34 M) for a second cross-linking step on gentle rotation at RT for 10 min.

Crosslinking reactions were immediately quenched by addition of 0.4 M glycine and incubated on ice for 5 min. Double-crosslinked cells were centrifuged (400 xg, 4°C, 5 min) then washed twice with 50 ml ice-cold PBS prior to re-suspension to $1x10^7$ cells/ml with 3 ml ice-cold buffer A. Cells were lysed on a rotating wheel at 4°C for 10 min and nuclei centrifuged (500 xg, 4°C, 5 min). The supernatant was discarded and pellets re-suspended to $1x10^7$ nuclei/ml in 3 ml ice-cold buffer B for incubation on a rotating wheel at 4°C for 10 min. Chromatin was pelleted by centrifugation (500 xg, 4°C, 5 min) and re-suspended to $2x10^7$ nuclei/ml 1.5 ml in ice-cold IP buffer I. 330 µl aliquots of chromatin were transferred to 1.5 ml Eppendorf tubes for sonication ('high' power setting, 30 second on/off cycles) using a Bioruptor™. The number of sonication cycles were determined empirically for each cell type to give a modal DNA fragment size of ~500 bp, and ranged from ~25 - 40 cycles.

Sonicated chromatin was cleared by centrifugation (16,000 xg, 4°C, 10 min) and the supernatant diluted with 2 volumes of IP buffer II. 100 µl of chromatin was reserved as an input control, and DNA isolated by addition of 0.2 M NaCl and 0.5 mg/ml Proteinase K and overnight incubation at 65°C then sequential phenol/phenol:chloroform/chloroform

extraction as described in 2.6.3. To confirm optimal sonication results, 10 µl of DNA input was analysed by electrophoresis on a 1.5% agarose gel with TAE/EtBr.

2.7.2 ChIP

10 µl of Dynabeads™-Protein G per IP were washed in 500 µl citrate phosphate buffer and stood on a DynaMag™-2 magnetic rack at RT until clear. The supernatants were discarded and beads re-suspended in 15 µl of citrate phosphate buffer with ChIP-grade antibody at the appropriate concentration (Table 2:4) and 0.5% BSA. Antibodies were pre-conjugated to beads on a rotating wheel at 4°C for 2 h. Pre-conjugated beads were washed in 500 µl citrate phosphate buffer to remove unbound antibody, then re-suspended in 15 µl citrate phosphate buffer/0.5% BSA. 500 µl of chromatin was added to the pre-conjugated beads and incubated on a rotating wheel at 4°C from 2 h - overnight (Table 2:4).

For FoxC1 ChIP-seq only, an additional pre-clearing step was performed prior to immunoprecipitation: 500 µl of chromatin was incubated with 10 µl of Dynabeads™-Protein G (without antibody) on a rotating wheel at 4°C for 1 h. Beads were isolated using a DynaMag™-2 magnetic rack and the supernatant added to α-FoxC1-conjugated beads as for other immunoprecipitations.

ChIP reactions were washed on a DynaMag™-2 magnetic rack using 500 µl volumes of buffers in the following order: wash buffer I (once); wash buffer II (twice); LiCl wash buffer (once); TE/NaCl wash buffer (twice). All wash steps were carried out on ice by gentle addition of the wash buffer and two 180° rotations of each sample tubes, still in the magnetic rack. Beads were re-suspended in 50 µl ChIP elution buffer and incubated on a ThermoMixer® C at 65°C and 500 rpm for 15 min. The first eluate was reserved and a second elution step was performed under the same conditions. Eluates were pooled

and 0.2 M NaCl and 0.5 mg/ml proteinase K added for reverse crosslinking at 65°C for > 4 h.

DNA was purified using AMPure® XP beads, using a 1.8X single-sided clean-up according to the manufacturer's instructions. ChIP efficiency was determined by qPCR as described in Quantitative PCR, using a standard curve prepared from sonicated chromatin input and specific primers against positive and negative control regions (Table 2:6). The fold-enrichment was calculated from the relative quantities of known/putatively positive-bound regions against control [negative] regions.

2.7.3 ChIP-seq library preparation

ChIP-seq libraries were prepared using a HyperPrep Kit for Illumina, according to the manufacturer's instructions. Amplified libraries were loaded with 1X green gel loading buffer onto on a 1.5% agarose gel with TAE/EtBr and electrophoresed at 80 V for 2.5 h. Library fragments of ~170 - 600 bp were isolated using a MinElute® Gel Extraction kit, according to the manufacturer's instructions, eluted in a final volume of 12 µl.

Library qPCR validation, quality control and quantitation were performed essentially as described in 2.6.5, using ChIP-specific primers detailed in Table 2:6.

2.7.4 ChIP-seq data processing

For RUNX1 and C/EBPα ChIP-seq data, raw reads were processed with Trimmomatic to remove low-quality reads and sequencing adapters. Reads were then mapped to the human genome assembly hg19 using Bowtie2 using the parameter -sensitive-local. For FoxC1 ChIP-seq, raw reads were instead trimmed using cutadapt[319] using -a ATATATATATATATATATAT as a parameter to remove artefactual repeats, then aligned using bowtie[320] to the hg19 genome with the following parameters: --all --best --strata -v

3 -m 1 –S, so as to obtain only uniquely aligned reads. Aligned reads were converted to BAM format using the SAMtools[321] view -S - -b -o command.

For all ChIP-seq analyses, coverage track generation and peak calling were performed using MACS using -g hs --keep-dup=auto -w -S as parameters. To remove artefactual peaks, peaks were intersected against the hg19 ENCODE blacklist[151] and UCSC simple repeats[322] using bedtools intersect with the -v switch.

2.7.5 ChIP-seq data analyses

To determine overlap of ChIP-seq binding enrichment between datasets, the MACS .bed output containing peak summits was converted to genomic ranges in ChIPpeakAnno[323], using the BED2RangedData function. ChIP peaks were considered to overlap if the summit points in either dataset were within 200 bp of each other. Overlaps between pairs of ChIP-seq datasets were visualised using the VennDiagram[324] package in R.

2.8 Electrophoretic Mobility Shift Assay (EMSA)

2.8.1 Preparation of nuclear extracts

$5x10^7$ cells were harvested by centrifugation (300 xg, RT, 5 min) and washed in 50 ml ice-cold PBS prior to re-centrifugation (300 xg, 4°C, 5 min). Cell pellets were re-suspended in 10 ml ice-cold buffer A and mechanically disrupted using a chilled 15 ml Dounce homogeniser on ice. Nuclei were pelleted by centrifugation (300 xg, 4°C, 5 min).

Nuclear pellets were re-suspended in 1/5 pellet volume of ice-cold buffer A, followed by very gradual addition of 5 pellet volumes of buffer C with thorough mixing. Nuclei were lysed by incubation on ice for 10 min with occasional shaking, and samples were cleared of insoluble chromatin by centrifugation (16,000 xg, 4°C, 10 min). Supernatants, containing soluble nuclear extracts, were transferred to pre-wetted Spectra/Por MYCO

3500 tubing for dialysis against buffer D at 4°C for 2 h. The protein concentration was determined by Bradford assay[325] using Bradford reagent and a spectrophotometer.

2.8.2 Preparation of EMSA probes

100 µM of sense and antisense PAGE-purified oligonucleotides (Table 2:6) were mixed then annealed by heating to 95°C for 10 min and gradual cooling to RT. 2 pmol of annealed duplexes were end-filled in 10.5 µl reactions containing 1X T4 polymerase buffer, 10 µM each of dGTP, dATP and dTTP, 0.5 units T4 DNA polymerase and 5 µCi of α-^{32}P dCTP. End-filling was carried out at RT for 10 min. α-^{32}P dCTP-labelled probes were purified using an Illustra ProbeQuant G-50 micro column, according to the manufacturer's instructions. Probes were eluted in 30 µl TE, then diluted to 320 µl (~0.005 pmol/µl, assuming 80% recovery) for direct use in EMSA binding reactions.

2.8.3 EMSA assays

A 4% polyacrylamide/0.5X TBE gel was prepared and allowed to set for at least 30 min. After this, the gel was pre-ran at 200 V for 1 h in 0.5X TBE, then allowed to cool prior to EMSA binding reactions.

Binding reactions comprised the following components added in order: 4 µg of poly(IC), 1X EMSA binding buffer, 0.5 mM DTT/PMSF, 1X EMSA Phosphatase Inhibitor Cocktail, 4 µg nuclear extract and 0.02 pmol of labelled EMSA probe. For competition assays, 2 pmol (i.e. a 100-fold excess relative to probe) of unlabelled EMSA duplex was added prior to nuclear extract and labelled probe.

Binding reactions were carried out at RT for 10 min, then loaded into the pre-ran gel and electrophoresed at 200 V for 90 min. Gels were fixed by immersion in 0.1% (w/v) CTAB/50 mM sodium acetate pH 5.5 for 45 min, then sandwiched between two sheets of Whatman® 3MM blotting paper for drying at 80°C for 1 h with a GD 2000 Gel Drier. Dried

gels were exposed to a cleared Storage Phosphor screen in an exposure cassette overnight, and images developed using a Personal Molecular Imager and Quantity One software.

2.9 Western blotting

2.9.1 SDS polyacrylamide gel electrophoresis (SDS-PAGE)

10 - 20 µg of nuclear extract of interest (prepared as in 2.8.1) were diluted to a volume of 10 µl with molecular biology-grade water. An equal volume of 2X Western blot sample buffer was added to each sample prior to boiling at 95°C for 10 min. Samples were loaded into separate wells of 4-20% Mini-PROTEAN TGX Stain-Free precast gels, together with 10 µl of Rainbow Molecular Weight Ladder for subsequent size estimation. Gels were electrophoresed in a vertical tank filled with 1X SDS Running Buffer at 15 mA for ~5 min, to allow proteins to enter the gel, then ran at 30 mA for 45 – 60 min, according to target protein sizes.

2.9.2 SDS-PAGE gel transfer, immunoblotting and visualisation of proteins

Proteins separated by SDS-PAGE were transferred onto nitrocellulose membranes using the 'Mixed MW' setting on a Trans-blot Turbo Transfer System. Membranes were blocked in 10 ml of Western blot blocking buffer at RT for 1 h, then transferred to containers with 10 ml of primary antibody in 5% milk/TBST (Table 2:4) for overnight incubation with gentle rocking at 4°C. The following day, membranes were washed with three successive 10 ml volumes of TBST with gentle rocking at RT, then transferred into a container containing 10 ml of secondary antibody in 5% milk/TBST for a 1 h incubation at RT with gentle rocking. Blots were washed three times as previously. Residual wash buffer was removed from membranes by carefully holding the blot against a clean paper towel to draw buffer off by capillary action.

Membranes were cut into two pieces to separate the experimental and internal loading control proteins prior to addition of a 1:1 mixture of Detection reagents 1 and 2 from an Amersham ECL Western Blotting Detection Reagent Kit, following the manufacturer's instructions. Membranes were briefly immersed in double distilled water prior to visualisation using a Gel Doc XR+ and Image Lab software.

2.10 Luciferase enhancer reporter gene assays

2.10.1 Generation of enhancer reporter plasmids

A subset of *FLT3*-ITD-specific DHSs identified in Cauchy *et al.* (2015) were amplified by overlap extension PCR from 10 ng of Jurkat genomic DNA using Phusion® High-Fidelity PCR Master mix, 500 nM of PCR cloning primers (Table 2:6), in 50 µl reactions in a T3000 thermocycler. Reactions were carried out using the following thermocycler parameters:

Table 2:9 **Thermocycling conditions for amplification of DHS fragments by PCR**

"x" denotes the optimal annealing temperature, specific to each primer pair. This was determined by first performing an initial optimisation PCR varying the temperature of x in a Primer Gradient Thermocycler and assessing target amplification by agarose gel electrophoresis.

Step	Temperature	Time	Number of cycles
Denaturing	98ºC	30 sec	1
Genomic amplification	98ºC	10 sec	5
	xºC	20 sec	
	72ºC	30 sec	
Adaptor amplicon amplification	98ºC	10 sec	25
	72ºC	30 sec	
Terminal elongation	72ºC	10 sec	1
	10ºC	∞	

Alternatively, equimolar quantities of an single-stranded oligonucleotides corresponding to three copies of the *FAM92A1* FOX:E composite site (Fox:Ex3) were annealed by heating to 95°C in a heatblock for 10 min then slowly cooling by removing the heatblock and placing on the bench. Cloning primers contained flanking sequences with BglII or BamHI/XhoI restriction sites to facilitate sticky-ended cloning into the parent vector, pXPG-TK229[326].

10 µl of each PCR product/annealed oligonucleotides were mixed with 1X purple gel loading buffer for electrophoresis on a 1.2% agarose gel with TAE/EtBr to confirm robust and specific target amplification. The remaining 40 µl of PCR products were cleaned up using a MinElute® PCR purification kit according to the manufacturer's guidelines, then digested in a 20 µl reaction with 20 U each of the appropriate restriction enzyme at 37°C for 3 h.

For each reporter construct to be prepared, 1 µg of pXPG-TK229 was digested with 20 U of restriction enzyme(s) in a 20 µl reaction at 37°C for 3 h. For constructs prepared via the BglII cloning strategy, 5' ends of the vector DNA were dephosphorylated in parallel by 1 U calf intestinal phosphatase.

Vector and insert DNA was purified from digest reactions using a QIAquick® PCR purification kit, according to the manufacturer's instructions, then quantified by 260 nm absorbance using a NanoDrop™ 2000 spectrophotometer. Ligations were carried out in 20 µl reactions comprising: 100 ng DNA (as a 3:1 molar ratio of insert:vector), 1X T4 DNA ligase buffer and 400 U T4 DNA ligase at RT for 1 h.

2.10.2 Transformation of *E. coli* and plasmid screening

50 µl of chemically-competent JM109 *E. coli* cells were transformed with 1 µl of ligation product by: incubation on ice for 20 min, heat shock at 42°C for 1 minute, recovery on ice

for 1 minute and incubation with 250 µl of SOC medium in an orbital shaking incubator (37°C, 250 rpm, 30 min). Transformed cells were plated onto LB agar plates with 100 µg/ml ampicillin (LB agar/amp) and placed in a static 37°C incubator overnight.

Colonies were picked to inoculate 2 ml cultures of LB medium with 100 µg/ml ampicillin (LB/amp), which were placed in a shaking incubator (37°C, 250 rpm, overnight). The following morning, 500 µl of each overnight culture was mixed with 500 µl of sterile 50% glycerol for long-term storage at -80°C. Plasmid mini-preps were made from the remaining 1.5 ml of overnight culture using a QIAprep® Miniprep kit, according to the manufacturer's instructions. Colonies were validated for correct insert sequence and orientation by Sanger sequencing (Source Bioscience) using the pXPG forward sequencing primer (Table 2:6).

2.10.3 Preparation of transfection-grade reporter plasmid stocks

Colonies verified to be positive for the desired plasmid were picked from master plates and re-streaked onto fresh LB agar/amp plates for propagation in a static 37°C incubator overnight. The following morning, a colony was picked to inoculate a 3 ml LB/amp starter culture in a shaking incubator (37°C, 250 rpm, 6 h). 100 µl of starter culture was introduced into a 200 ml LB/amp culture for larger-scale culture (37°C, 250 rpm, overnight).

The following morning, bacteria were transferred to 225 ml Falcon tubes for collection by benchtop centrifugation (2630 xg, 4°C, 15 min). The supernatants were discarded and the pellets detached from the bottom of Falcon tubes by gentle vortexing. Pellets were resuspended in 8 ml of Maxi-prep Resuspension Buffer on ice, followed by addition of 15 ml of Maxi-prep Alkaline Lysis Buffer. Samples were mixed by gentle inversion then stood on ice for 3 min. 10 ml of Maxi-prep Neutralisation Buffer was added to each

sample and samples gently inverted until evenly mixed. Samples were stood on ice for 10 min then re-centrifuged (2630 xg, 4°C, 10 min).

The supernatants were transferred to 50 ml Oakridge tubes, avoiding transfer of debris and bubbles, then centrifuged (20,000 rpm [Beckman JA-25 rotor], 4°C, 15 min). 28 ml of supernatant was removed and transferred to a 50 ml Falcon, to which 17 ml of isopropanol was added. These tubes were incubated at RT for 15 min then centrifuged (2,000 xg, RT, 20 min). The supernatants were discarded and pellets washed by vortexing in 20 ml of 70% ethanol, then re-centrifuged (2,000 xg, RT, 3 min). All traces of residual ethanol were removed using a pipette. Pellets were re-suspended in 2.6 ml of 1X TE for 2 hours under vigorous shaking on an orbital platform. At this point, 2.9 g of CsCl were added to each tube, dissolved fully, followed by 120 µl of 10 mg/ml EtBr. Following addition of EtBr, tubes were protected from light by covering with foil when on the bench. Tubes were centrifuged (2,000 xg, RT, 5 min) and the contents transferred to 3.3 ml Optiseal TLN100 tubes. Pairs of tubes were balanced to within 0.01 g of each other by addition of additional CsCl/TE solution. Samples were centrifuged at ultra-high speed (85,000 rpm [Beckman TLN-100 rotor], RT, 18 h, brake = 0).

The following day, the lower visible band in the tube (comprising partially-purified plasmid DNA) was harvested from Optiseal tubes using a 1 ml syringe and 23 gauge needle and transferred to a fresh 3.3 ml Optiseal tube. The tubes were filled and made to within 0.01 g of each other by addition of CsCl/TE solution with additional 1 mg/ml EtBr. Tubes were re-centrifuged (95,000 rpm [Beckman TLN.100 rotor], RT, 5 h, brake = 0) and high purity plasmid DNA harvested as described previously.

Harvested plasmid DNA was transferred to 1.5 ml Eppendorf tubes and 200 µl of HyClone™ water added. An equal volume of *n*-butanol was added to each tube and

mixed vigorously. The lower aqueous phase was transferred to a fresh 1.5 ml Eppendorf tube using a pipette and this process was repeated a further six times. After seven *n*-butanol extractions, samples were transferred to 15 ml Falcon tubes and made up to a volume of 2 ml with 50 mM NaAc. To this, 4 ml of 100% ethanol was added and mixed gently by rocking. Samples were incubated at -20°C for 10 min then centrifuged (2000 xg, 4°C, 15 min). The supernatants were discarded and all traces of ethanol removed prior to the addition of 400 µl of 1X TE. Samples were incubated (37°C, 25 min) with occasional mixing, until the DNA pellets were fully dissolved.

Dissolved plasmid DNA was transferred into 1.5 ml Eppendorf tubes, to which 20 µl of 3 M NaAc and 900 µl 100% ethanol were added. Samples were gently mixed until an obvious fibrous precipitate of DNA appeared, then stood at RT for 5 min. DNA was pelleted by centrifugation (18500 xg, 4°C, 2 min), the supernatant discarded and the pellet washed in 500 µl of 70% ethanol. Washed pellets were re-centrifuged as before, supernatant removed and the pellets were left in a sterile tissue culture hood until dried, typically for ~20-30 min. At this stage, highly purified plasmid DNA was resuspended in 500 µl of 1X TE and incubated at 37°C for 30 min to permit rapid dissolution. Samples were mixed gently and stood overnight prior to quantitation the following day by 260 nm absorbance using a NanoDrop™ 2000.

200 ng of purified plasmid were electrophoresed on a 0.8% agarose gel with TAE/EtBr, to confirm purification of high-quality, supercoiled plasmid.

2.10.4 Dual Luciferase Transfection assay

All reporter plasmids to be tested were re-quantitated prior to each experiment to ensure accurate amounts were used. FUJIOKA cells in exponential growth were counted and harvested by centrifugation (x300 g, RT, 5 min) and resuspended in media to give a cell

density of 11.25 x 10^6/ml. 400 µl of FUJIOKA cells (i.e. 4.5 x 10^6) were transferred to 0.4 mm electroporation cuvettes. 5 µg of experimental plasmid (using the pXPG-TK229 backbone) and 1 µg of normalisation control plasmid (pRL-TK750) were added to each cuvette and mixed gently, then incubated at RT for 10 min. Cells were electroporated (270 V, 950 µF, exponential decay) using a Gene Pulser Xcell™ electroporator system, then allowed to recover (RT, 5 min). Following this, transfected FUJIOKA cells were carefully transferred to 6 well plates containing 9.6 ml of media and returned to routine culture conditions.

The following day, cells to be stimulated were treated as described in 2.1.2. Following stimulation, all cells were harvested (x300 g, RT, 5 min) and prepared for luciferase reporter assays using a Dual-Luciferase Reporter Assay System kit, according to the manufacturer's instructions. Luciferase data were acquired using the 'Dual-Luciferase' setting of a GloMax®-Multi+ Microplate Multimode Reader. Normalised luciferase data were calculated by dividing the Firefly (experimental) luciferase signal by the internal Renilla normalisation control signal.

2.11 CRISPR deletion of the *FoxC1* locus

2.11.1 CRISPR sgRNA Design

sgRNA sequences were identified 3' to the *FoxC1* start codon and 3' to the stop codon using the Zhang group online CRISPR design tool. Guide sequences computationally predicted by the design tool to have minimal off-target effects were selected, and the 5' ends of each were truncated by 2 nt according to Fu *et al* (2014)[327] to further improve cleavage specificity. These sequences were ordered as pairs of complementary oligonucleotides to be compatable with downstream cloning steps (Table 2:6).

2.11.2 Generation of CRISPR-Cas9 expression constructs

100 µM of sense and antisense oligonucleotides were phosphorylated and annealed in parallel using 5 U of T4 Polynucleotide Kinase (New England Biolabs) and 1X T4 Ligation Buffer (New England Biolabs) by incubation at 37°C for 30 min, 95°C for 5 min, before slowly cooling to RT.

Annealed oligonucleotides were cloned into sSpCas9(BB)-2A-GFP using the Golden Gate assembly cloning strategy[328] as follows: 100 ng of circular PX458, 1 µM of annealed oligonucleotide, 1X NEB Buffer 2.1, 20 U BbsI restriction enzyme, 10 mM ATP, 5 µg BSA, 750 U T4 DNA ligase in a 50 µl reaction volume. Reactions were carried out in a T3000 thermocycler using the following programme: 20 cycles of 37°C for 5 min, 20°C for 5 min; 1 cycle of 80°C for 20 min.

Chemically competent JM109 *E. coli* cells were transformed and colonies screened as described in 2.10.1, using the U6 promoter forward primer (Table 2:6) to verify positive clones by Sanger sequencing (Source Bioscience).

Colonies verified by sequencing to contain the correct sgRNA insert sequence were used to inoculate a 2 ml starter culture of LB/amp (37°C, 250 rpm, 6 h). 100 µl of starter culture was used to inoculate 200 ml of LB/amp in an Erlenmeyer flask for maxi-prep culture (37°C, 250 rpm, overnight). Endotoxin-free plasmid maxi-preps were obtained using a NucleoBond® Xtra Midi EF kit, following the manufacturer's instructions.

2.11.3 Transfection of CRISPR reagents

$4.5x10^6$ FUJIOKA cells were suspended in 400 µl of antibiotic-free RPMI and 10% heat-inactivated foetal calf serum for transfection with either 5 µg of PX458 (WT) or 5 µg of pairs of constructs containing 5' and 3' sgRNAs for *FOXC1* gene deletion. Cells were transferred to 0.4 cm gap width cuvettes for electroporation (325 V, 30 msec, 0.4 cm)

using a Gene Pulser Xcell™ electroporator system (Bio-Rad). Cells were allowed to recover at RT for 10 min, then transferred to 6 well plates in a 10 ml final volume of 40% Fujioka-conditioned medium. 48 h post-transfection, single GFP-positive cells were sorted using a FACSAria™ II cell sorter into 96 well plates containing 100 µl of 40% Fujioka cell-conditioned media (2.1.3) per well. Cells were expanded for at least 14 days prior to screening for deletions.

2.11.4 PCR screening of clones for deletions

Cells from each clone were isolated by centrifugation (300 xg, RT, 5 min) and genomic DNA extracted as described in 2.6.3. 20 ng of genomic DNA per clone was screened using Phusion® High-Fidelity PCR Master mix, 1X GC buffer, 1 M betaine and 200 nM of PCR cloning primers designed against either side of the intended *FOXC1* deletion sites (Table 2:6), in 20 µl reactions in a T3000 thermocycler.

Table 1:1 **Thermocycling conditions used for PCR screening of CRISPR clones**

Step	Temperature	Time	Number of cycles
Denaturing	98ºC	30 sec	1
Genomic amplification	98ºC	10 sec	25
	62ºC	20 sec	
	72ºC	60 sec	
Terminal elongation	72ºC	5 min	1
	10ºC	∞	

Clones carrying deletions were cultured under normal conditions to give sufficient numbers for analyses of *FOXC1* gene expression (2.5.3).

2.12 shRNA-mediated knockdown of *FOXC1* expression

2.12.1 Cloning of shRNA expression vectors

shRNA sequences targeting the human *FOXC1* transcript were designed using the Broad Institute's Genetic Perturbation Platform and selected based on the software algorithm's highest predicted knockdown scores. These sequences were modified to have HpaI and XhoI restriction sites compatible for cloning into the LeGO-iG vector immediately downstream of the U6 promoter (Table 2:5 and Table 2:6).

1 µg of sense and antisense oligonucleotides corresponding to these sequences (table) were annealed by heating to 95°C in a heatblock for 5 min then cooling slowly to RT by placing the heatblock on the bench. Annealed oligonucleotides were then digested using 10 units each of HpaI and XhoI (37°C, 4 h). 5 µg of empty LeGO-iG vector was digested under the same conditions, then dephosphorylated by addition of 1 unit of Calf Intestinal Phosphatase and incubation (37°C, 30 min). Digested LeGO-iG vector was purified by 1% agarose gel electrophoresis and a MinElute® Gel Extraction Kit, according to the manufacturer's instructions.

Digested oligonucleotide duplexes and vector backbone were ligated in a 3:1 insert:vector ratio using T4 DNA ligase (37°C, 2h), according to the manufacturer's instructions. JM109 *E. coli* were transformed with ligation products as previously described (2.10.2) prior to plating onto LB agar/amp plates before incubation (37°C, overnight). The following morning, plates were put into short-term storage at 4°C. That evening, single colonies were picked to inoculate 2 ml cultures of Terrific broth with 100 µg/ml carbenicillin (TB/carb) and placed in a shaking incubator (37°C, 250 rpm, overnight). The following morning, 500 µl of each overnight culture was mixed with 500 µl of sterile 50% glycerol for long-term storage at -80°C. Plasmid mini-preps were made from the remaining 1.5 ml of overnight culture using a QIAprep® Miniprep kit, according

to the manufacturer's instructions. Colonies were validated for correct insert sequence and orientation by Sanger sequencing (Source Bioscience) using the U6 forward sequencing primer (Table 2:6).

Colonies verified to be positive for the intended sequence were picked to inoculate a 3 ml TB/carb starter culture in a shaking incubator (37°C, 250 rpm, 6 h). 100 µl of starter culture was introduced into a 300 ml TB/carb culture for larger-scale culture (37°C, 250 rpm, overnight). The following morning, endotoxin-free maxi-preps were prepared using a NucleoBond® Xtra Midi EF kit, according to the manufacturer's instructions. Plasmids were quantitated by A260 absorbance using a NanoDrop™2000 spectrophotometer.

2.12.2 Production of lentivirus in HEK 293T cells

HEK 293T cells were re-plated 24 h prior to transfection, to reach a confluency of 80-90% at time of transfection. On the day of transfection, fresh DMEM supplemented with 10% FCS, 2 mM glutamine and penicillin/streptomycin was exchanged with the previous media present on HEK 293T culture plates. Prior to transfection, TransIT®-293 was brought to RT. For each transfection, a DNA mix was prepared containing the shRNA expression vector (30 µg) and the packaging vectors: Gag/Pol (1.2 µg); Tat (1.2 µg); Rev (1.2 µg) and VSV-G (2.4 µg). For each small ($15cm^3$) TC dish to be transfected, 2 ml of OptiMEM™ was mixed with 90 µl of TransIT®-293 and incubated (RT, 15 min). The DNA mix described was then added to the OptiMEM™:TransIT®-293 mixture and incubated further (RT, 15 min). After this, DNA:TransIT®-293 mixture was added dropwise to the HEK 293T plates. Viral supernatant was collected after 24 h and every 12 h thereafter for a total of 48 h.

2.12.3 Concentration of Lentiviral Supernatants

Viral supernatant collected from the previous step was centrifuged at (1,660 xg, 4°C, 15 min) to pellet cell debris. Cleared supernatant was then passed through a 0.45µm disc

filter to remove residual cells and concentrated using a Centricon® Plus-70 100kDa filter (Millipore). Prior to loading viral supernatant, the column was washed with sterile water then centrifuged (2000 xg, 4°C, 25 min). Viral supernatant was then concentrated according to the manufacturer's instructions.

2.12.4 Lentiviral transduction of FUJIOKA cells

$2x10^6$ FUJIOKA cells were transduced with the lentiviral concentrates prepared from 30 cm^3 of transduced HEK 293T cells and 8 µg/ml of polybrene by spinoculation (1500 xg, 32°C, 2 h) in non-tissue culture treated 6-well plates. Following spinoculation, cells were allowed to recover in an incubator under normal culture conditions overnight. The following morning, viral media was aspirated and exchanged with fresh media. 72 hours post-transduction, the efficiency of transduction was estimated by measuring the proportion of eGFP-positive cells using flow cytometry. Cell sorting for GFP positive cells was performed using a FACSaria II by the College of Medical and Dental Sciences Core Technology Cell Sorting facility.

2.12.5 Growth analyses of transduced FUJIOKA cells

Sorted GFP+ FUJIOKA cells were seeded at a concentration of $0.1x10^6$/ml in 10 ml of media and returned to the incubator. Cells counts were taken every 3-4 days then passaged to $0.1x10^6$/ml in 10 ml of media. Cell growth rates were determined by calculating the true cell number (assuming all cells were retained following each passage) and doubling times.

Chapter 3: FoxC1 co-operates with Hoxa9 to activate a specific chromatin and transcriptional signature in AML

A major focus of this study was to search for epigenetic and transcriptional mechanisms underlying the leukaemogenic consequences of FoxC1 derepression in AML. In this chapter I will explore mechanisms whereby FoxC1 becomes activated in parallel with HOX family genes and cooperates with Hoxa9 to activate specific genes and promote the development of AML, focussing on a mouse model developed by Tim Somervaille's laboratory.

3.1 *FOXC1* expression is associated with enhanced accessibility of the *HOXA* cluster in human primary AML patients and cell lines

Given that *FOXC1* mRNA expression is strongly correlated with *HOXA9* expression in primary human AMLs[300], we sought to investigate whether changes in *FOXC1* expression were associated with specific changes in the accessible chromatin landscape at the *HOXA9* locus. To this end, we initially performed DNase-seq and RNA-seq on cells from a representative *FOXC1*^{high} FLT3-ITD primary human AML sample and the FUJIOKA cell line, which features comparatively high expression of *FOXC1* relative to other human AML cell lines (Figure 3:1). These data were compared with data from cells expressing low or modest levels of *FOXC1*, including a representative FLT3-WT primary AML sample (both CD34$^+$ and CD34$^-$ populations), K562 cells, and normal CD34$^+$ PBSCs featuring high expression levels of *HOX* genes including *HOXA9*[151,271,329].

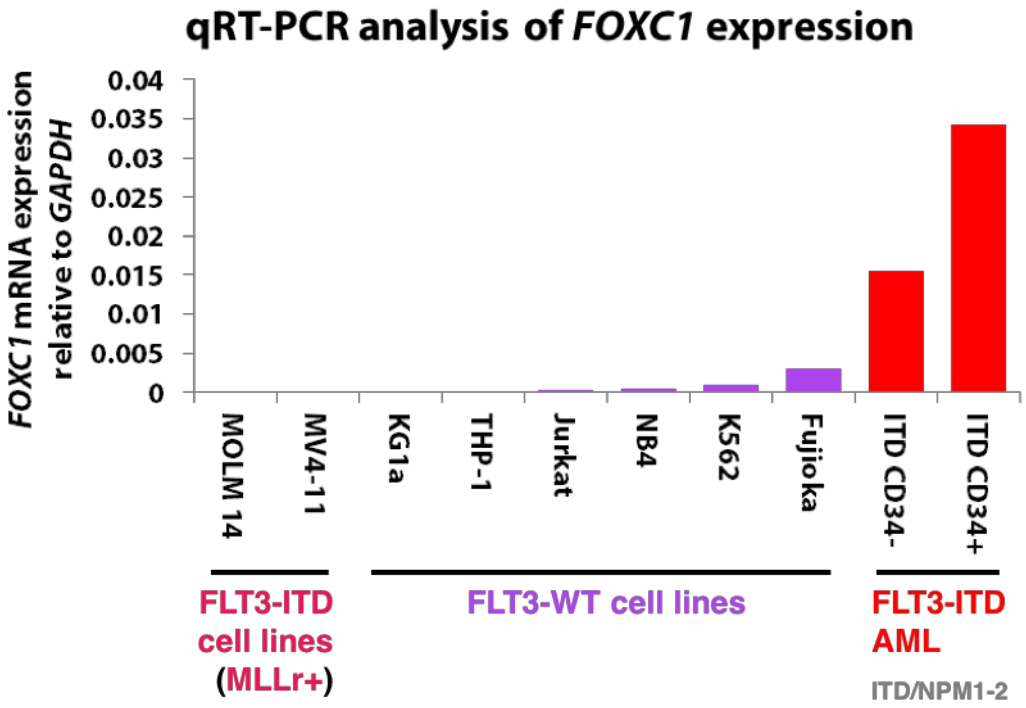


Figure 3:1 ***FOXC1* is substantially expressed in primary FLT3-ITD AML patient samples, but not cell lines with FLT3-ITD.**

Expression levels of FOXC1 mRNA were compared by qRT-PCR analyses of cDNA prepared from a panel of leukaemia cell lines with or without FLT3-ITD, and the CD34 enriched/negative fractions (CD34+ and CD34-, respectively) of a representative FLT3-ITD primary human AML sample, ITD/NPM1-2.

N.B. both FLT3-ITD AML cell lines are driven by MLL rearrangement oncogenes (MLLr+): MOLM-14 features the MLL-AF9 fusion oncogene, while MV4-11 carries the MLL-AF4 translocation.

Firstly, we confirmed that robust expression of *FOXC1* in primary FLT3-ITD AML cells and FUJIOKA cells correlated with increased chromatin accessibility at the *FOXC1* promoter-proximal region and gene body, as shown by DNase-seq analyses (Figure 3:2). By contrast, a representative primary FLT3-WT AML sample and $CD34^+$ PBSCs were all substantially less hypersensitive at the *FOXC1* promoter. However, despite only modest detectable expression of *FOXC1* in K562 cells by qRT-PCR analyses, the promoter-proximal region of *FOXC1* was found to be hypersensitive in contrast to the gene body, which had little to no detectable DNase-seq signal.

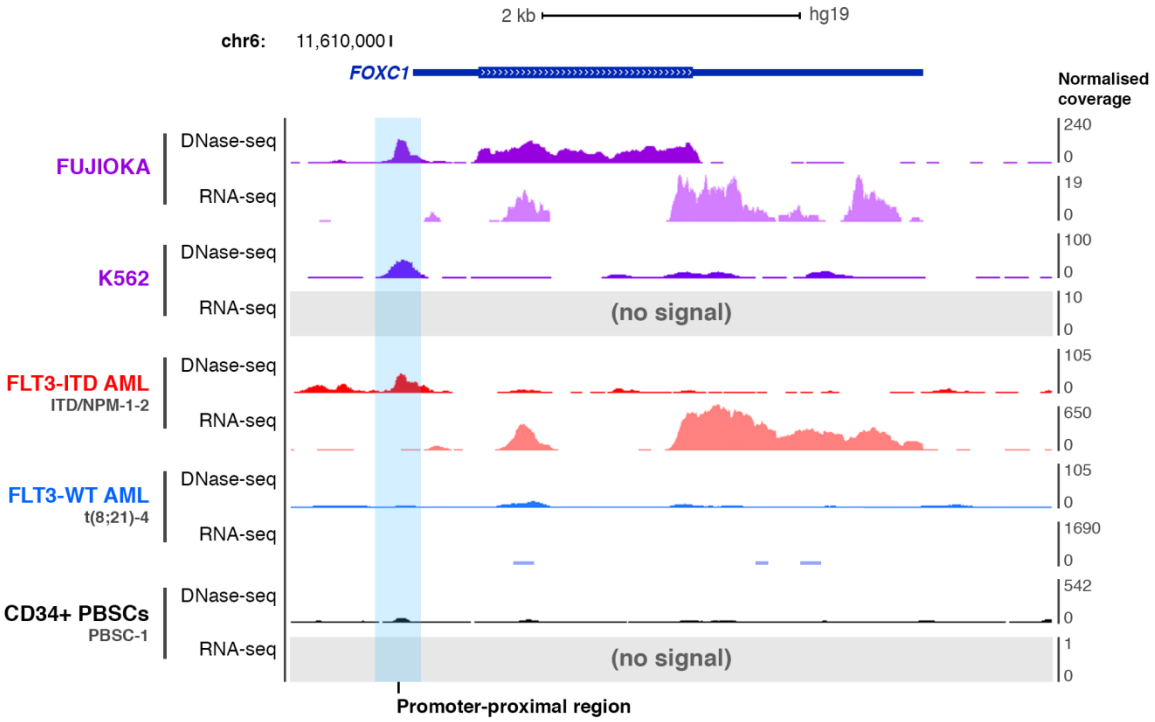


Figure 3:2 **The *FOXC1* locus is accessible and transcribed in FLT3-ITD AML patients and the FUJIOKA cell line.**

UCSC genome browser view of the human FOXC1 locus, showing DNase-seq and RNA-seq data for the FoxC1^{high} FUJIOKA cells and FoxC1^{low} K562 cells, together with data from representative FLT3-ITD and FLT3 wild-type (FLT3-WT) primary AML samples and normal control CD34+ PBSCs. The promoter-proximal region of FOXC1 is highlighted in blue.

At the *HOXA* gene cluster (Figure 3:3), a broad increase in chromatin accessibility and *HOXA* gene expression across the entire ~100 kb chromatin domain was associated with elevated expression of *FOXC1* in both primary human AML samples and FUJIOKA cells. These included at least seven DHSs that were present in the primary FLT3-ITD AML and FUJIOKA cells but not the representative FLT3-WT AML sample or K562 cells. However, manual inspection of these DHSs revealed an absence of the 5'-RYMAAYA-3' FOX consensus binding motif in all but one of these DHSs (DHS 3) that may indicate direct regulation of the *HOXA* cluster by FoxC1.

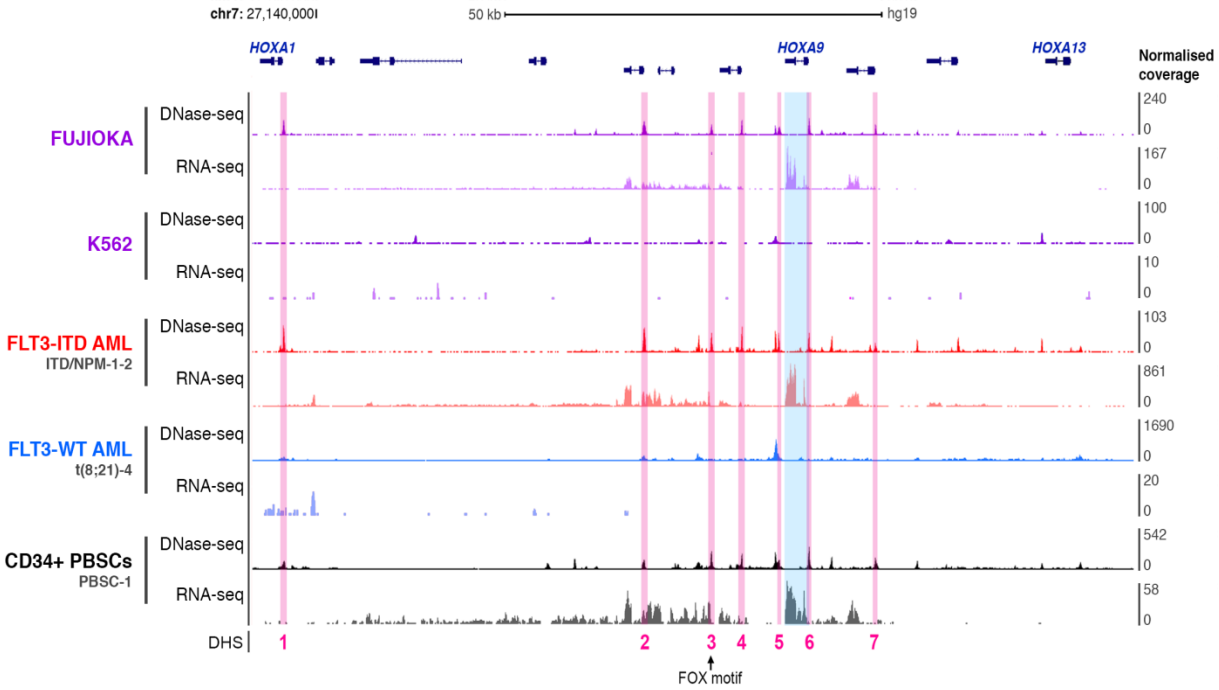


Figure 3:3 **The *HOXA* gene cluster is also accessible and actively expressed in leukaemia cell lines and patients expressing *FOXC1*.**

UCSC genome browser view of the human *HOXA* gene cluster, showing DNase-seq and RNA-seq data for the FoxC1^{high} FUJIOKA cells and FoxC1^{low} K562 cells, together with data from representative FLT3-ITD and FLT3 primary AML samples and normal control CD34^{+} PBSCs. The *HOXA9* gene is highlighted in blue. DHSs specific to FoxC1^{high} samples are highlighted in pink and numbered at the bottom.

Having substantiated a previously described correlation between FoxC1 and HOXA9 expression at the level of chromatin accessibility, we went on to perform a detailed molecular analysis of open chromatin, TF occupancy and gene expression in a Hoxa9+FoxC1-dependent mouse model of AML.

3.2 FACS immunophenotyping confirms robust engraftment of donor leukaemia cells, abnormal myeloid bias and impaired B cell differentiation in Hoxa9+FoxC1 AMLs

A significant aspect of studies published by Somerville *et al.* (2015) was the development of a mouse model of AML driven by the ectopic expression of human *FOXC1* together with murine *Hoxa9*. As summarised in Figure 3:4, this mouse model was generated from donor BM cells transduced with retroviral expression vectors for *Hoxa9* with empty

mouse mammary tumour virus (MMTV) vector, or in combination with either *FOXC1* or *Meis1*, an extensively-documented leukaemogenic partner of HOX proteins[330–333].

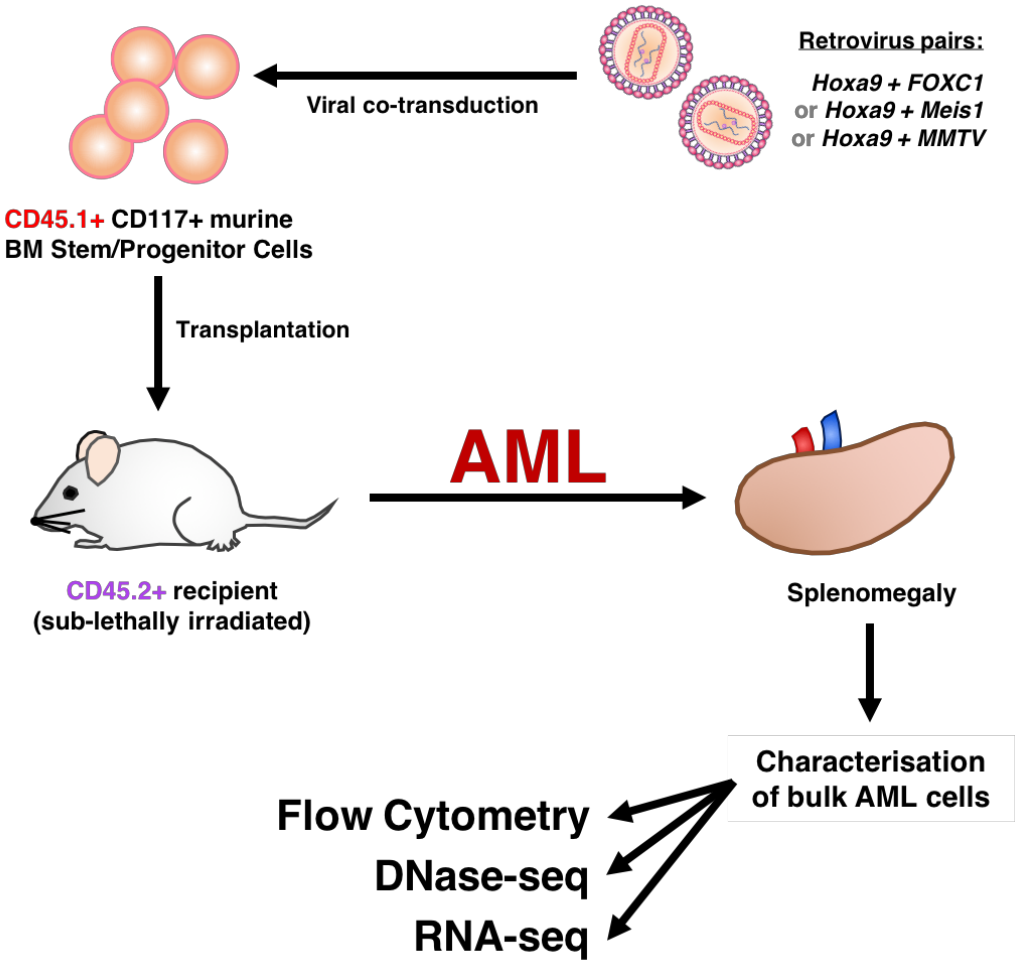


Figure 3:4 **Overview of the Somervaille Hoxa9+FoxC1-dependent mouse AML model and downstream molecular analyses conducted in this study.**

Cryopreserved neoplastic/AML samples from the spleens of moribund mice generated in Somerville *et al.* (2015) were subjected to phenotypic analyses by multicolour flow cytometry, followed by DNase-seq and RNA-seq to identify changes in accessible chromatin and gene expression.

Previous studies of this mouse model involved extensive characterisation of the phenotypic consequences arising from co-expression of these proto-oncogenes, but the *cis*-regulatory elements and target genes through which FoxC1 and Hoxa9 act to promote leukaemogenesis remained unknown. To address this gap in knowledge, we

established a collaboration with the Somervaille laboratory to perform genome-wide chromatin accessibility, genomic occupancy and gene expression profiling on cryopreserved spleen samples from these leukaemic mice. To ensure that the spleen cells under investigation were predominantly engrafted AML cells, the purity of leukaemic cells within these samples was verified by flow cytometry prior to any downstream analysis.

First, we assessed the extent to which CD45.1$^+$ donor cells engrafted into the recipient animals from which the spleen samples were prepared (Figure 3:5). These analyses revealed robust engraftment of CD45.1$^+$ donor cells in samples from all three Hoxa9 co-expression mouse models, with Hoxa9+FoxC1 recipient mice featuring the highest proportion of donor cells (90.3% in a representative sample). CD45.1$^+$ donor cell engraftment in the Hoxa9+Meis1 and Hoxa9+MMTV recipients was extensive, but not as high as observed in Hoxa9+FoxC1 recipients. Typical engraftment populations for these mouse represented ~75.6% and 63.9% of all cells in samples from Hoxa9+Meis1 or Hoxa9+MMTV recipients, respectively. Next, we assessed myeloid lineage bias within the CD45.1$^+$ donor fractions of cells by examining expression of the CD11b and Gr1 myeloid surface markers. Our analyses revealed that the CD45.1$^+$ fractions from all mouse models were overwhelmingly enriched for the CD11b$^+$Gr1$^+$ myeloid markers, ranging from 81.7% to 97.8% of double-positive cells. In the case of Hoxa9+FoxC1 recipients, there was a near-total absence of any cells lacking these markers (97.8% CD11b$^+$Gr1$^+$), indicating a strong myeloid bias in the engrafted donor cell populations within recipient spleens. Furthermore, the CD11b$^+$Gr1$^+$ fraction of Hoxa9+FoxC1 donor cells almost completely lacked expression of the B cell lineage marker B220 (97.0% B220$^-$Gr1$^+$CD11b$^+$). The overwhelming majority of Hoxa9+FoxC1 donor cells instead expressed the immature B cell precursor marker CD43 (96.7% CD43$^+$Gr1$^+$CD11b$^+$),

which is not consistent with normal B cell differentiation. This contrasted with Hoxa9+Meis1 and Hoxa9+MMTV CD45.1^{+} donor cells, which both featured modestly higher surface expression of B220 (1.1% and 3.9% for Hoxa9+MMTV and Hoxa9+Meis1 respectively).

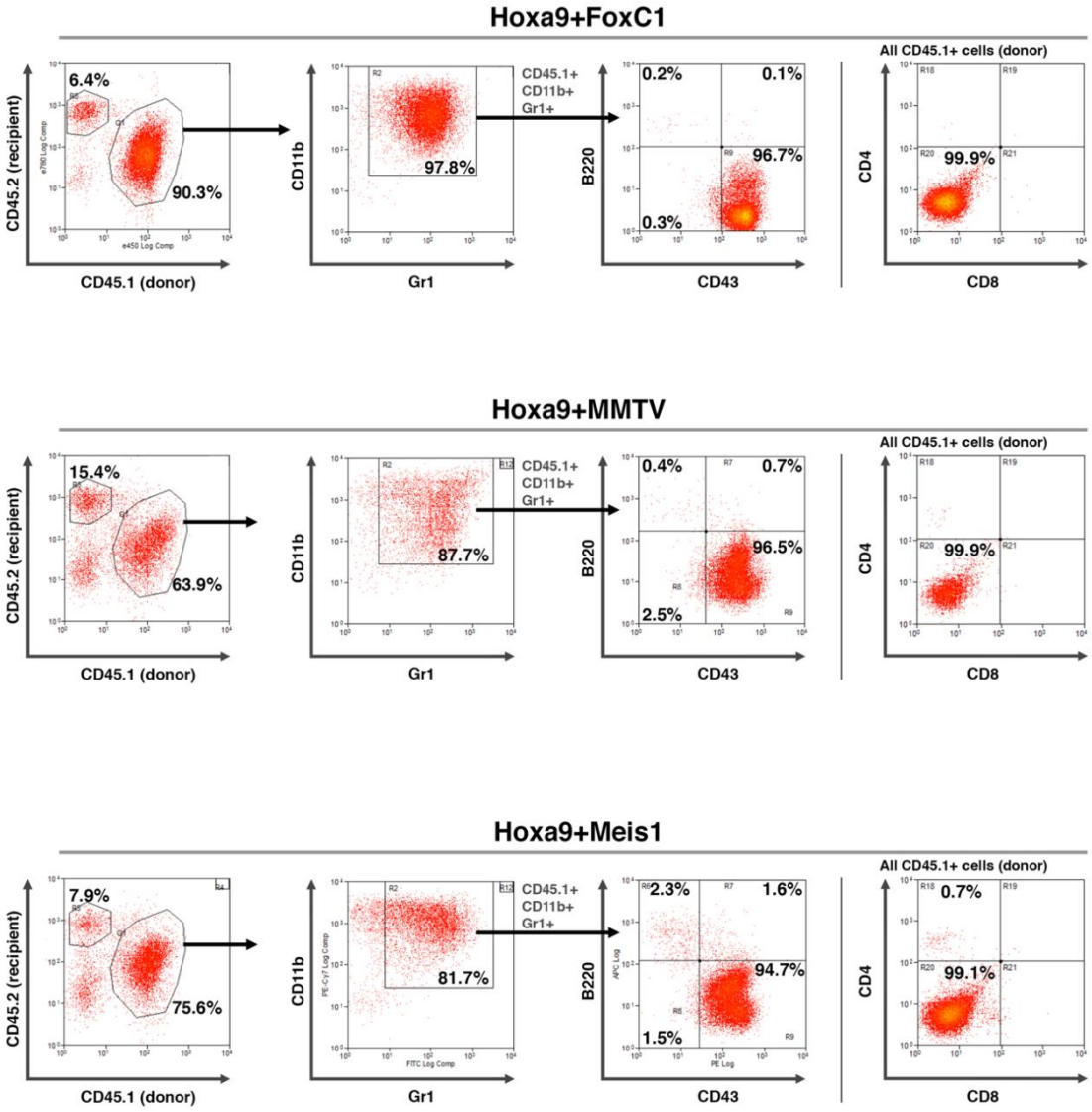


Figure 3:5 **Hoxa9+FoxC1 recipient mice feature substantial splenic infiltration of leukaemic cells as measured by FACS.**

The proportion of engrafted neoplastic cells in cyropreserved spleen samples was determined as the proportion of CD45.1+ donor cells within bulk samples also containing CD45.2+ host cells. Myeloid lineage bias was ascertained from the proportion of neoplastic donor cells expressing both CD11b and Gr1 surface markers. Evidence of B or T cell differentiation was examined by measuring the surface expression of the immature B lineage markers CD43 and B220, or the T cell markers CD4 and CD8.

Finally, CD45.1$^+$ donor cell fractions from all samples lacked expression of the T cell lineage markers CD4 and CD8. Together with the near-complete effacement of CD45.2$^+$ recipient cells within Hoxa9+FoxC1 samples (6.4% CD45.2$^+$), these observations confirmed an abnormal myeloid lineage bias in spleens harbouring Hoxa9+FoxC1 AML cells.

In contrast, FACS analyses of the minor populations of contaminating CD45.2+ host cells (Figure 3:6) revealed comparatively lower surface expression of the myeloid markers CD11b and Gr1 (20.1% to 37.1% Gr1$^+$CD11b$^+$). Furthermore, CD45.2$^+$ host cells for all samples featured higher, more variable surface expression of the immature B cell marker B220, ranging from 13.7% to 51.7%. We also observed a markedly greater abundance of normal B lineage precursors, including pre-pro B cells as defined by the CD43$^+$B220$^+$ immunophenotype (3.5% to 7.4%) and CD43$^-$B220$^+$ pro-B cells (10.2% to 53.4%), in all three mouse models, consistent with more normal B cell differentiation in the host cell compartment. As with the CD45.1$^+$ donor cells for each of the three different models, all CD45.2$^+$ host cells were devoid of CD4 expression, a marker used to broadly define all four of the most common subsets of T helper cells[334]. However, there was a greater abundance of host CD4$^-$CD8$^+$ cytotoxic T cells as compared to the CD45.1$^+$ donor cell fractions from all mice.

To summarise, the relative proportions of different sub-populations within the CD45.2$^+$ host cell compartment – particularly those with B lymphoid markers – were more consistent with that of a normal spleen. However, these cells formed a much smaller proportion of the overall number of cells in these samples compared to the CD45.1$^+$ donor cell fractions. Collectively, our FACS analyses support previous observations of BM from the same mice that indicated a disruption of residual host haematopoiesis and the robust engraftment of abnormal myeloid donor cells[300].

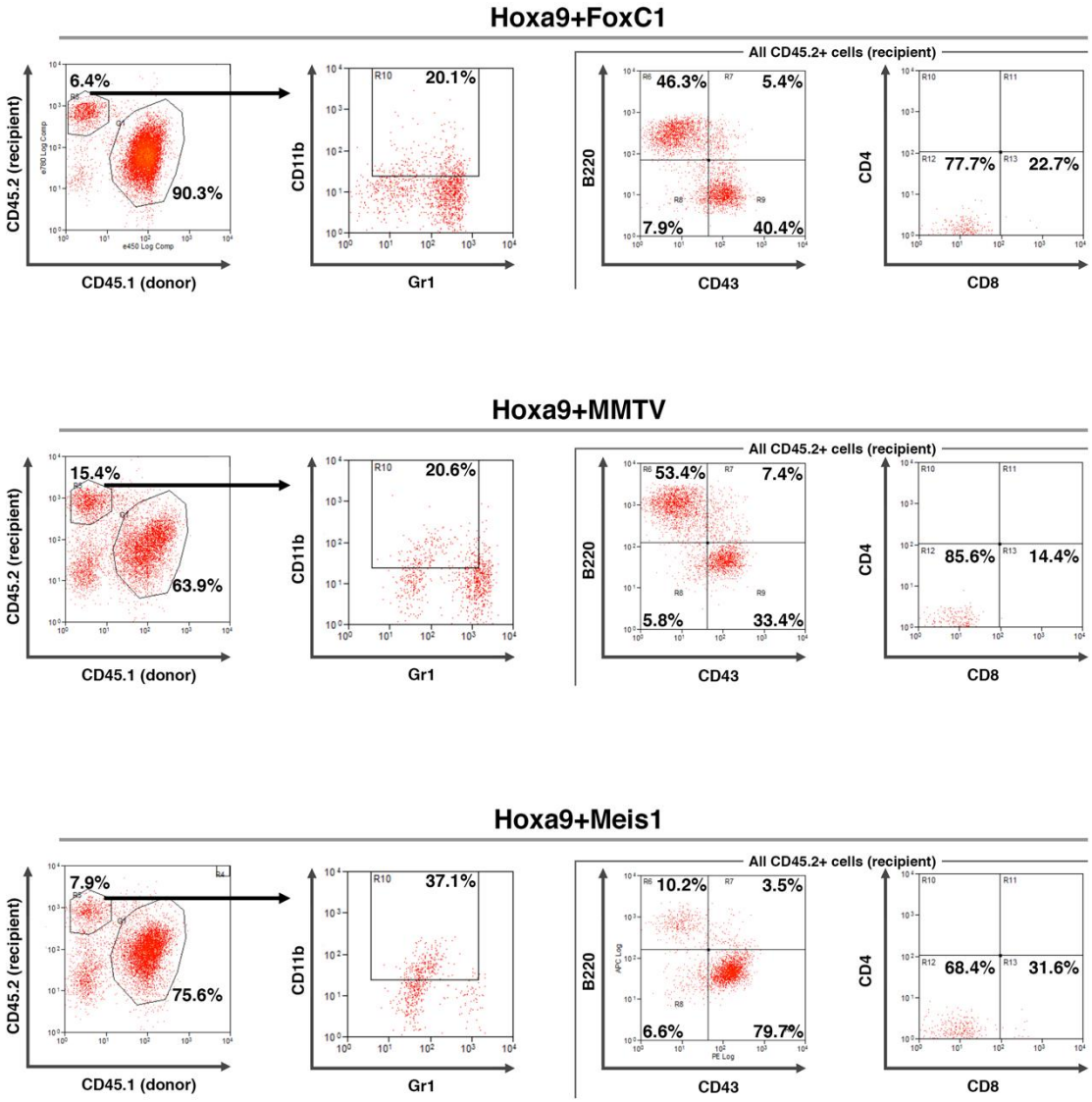


Figure 3:6 **Minor populations of contaminating host cells exhibit more mature B and T lineage characteristics as measured by FACS**

The proportion of recipient host cells in cyropreserved spleen samples was determined as the proportion of cells carrying CD45.2 expression. Myeloid, B and T lymphoid lineage characteristics using the same combination of immunophenotypic markers as for donor leukaemia cells described in Figure 3:5 on all cells with surface expression of CD45.2.

3.3 FoxC1 collaborates with Hoxa9 to activate a specific chromatin signature in AML

To gain insights into changes in the accessible chromatin landscape arising from Hoxa9 and FoxC1-dependent leukaemic transformation, we performed genome-wide DNase-seq on the same bulk splenic cells that had been profiled by FACS. For each of the three mouse models studied, several biological replicates were used to increase the robustness and reliability of analyses. In each case, one representative sample for each mouse model was taken forward for high depth sequencing to serve as an anchor sample used to generate the highest quality data (defined here as Fox-1, MTV-1 and Meis-1).

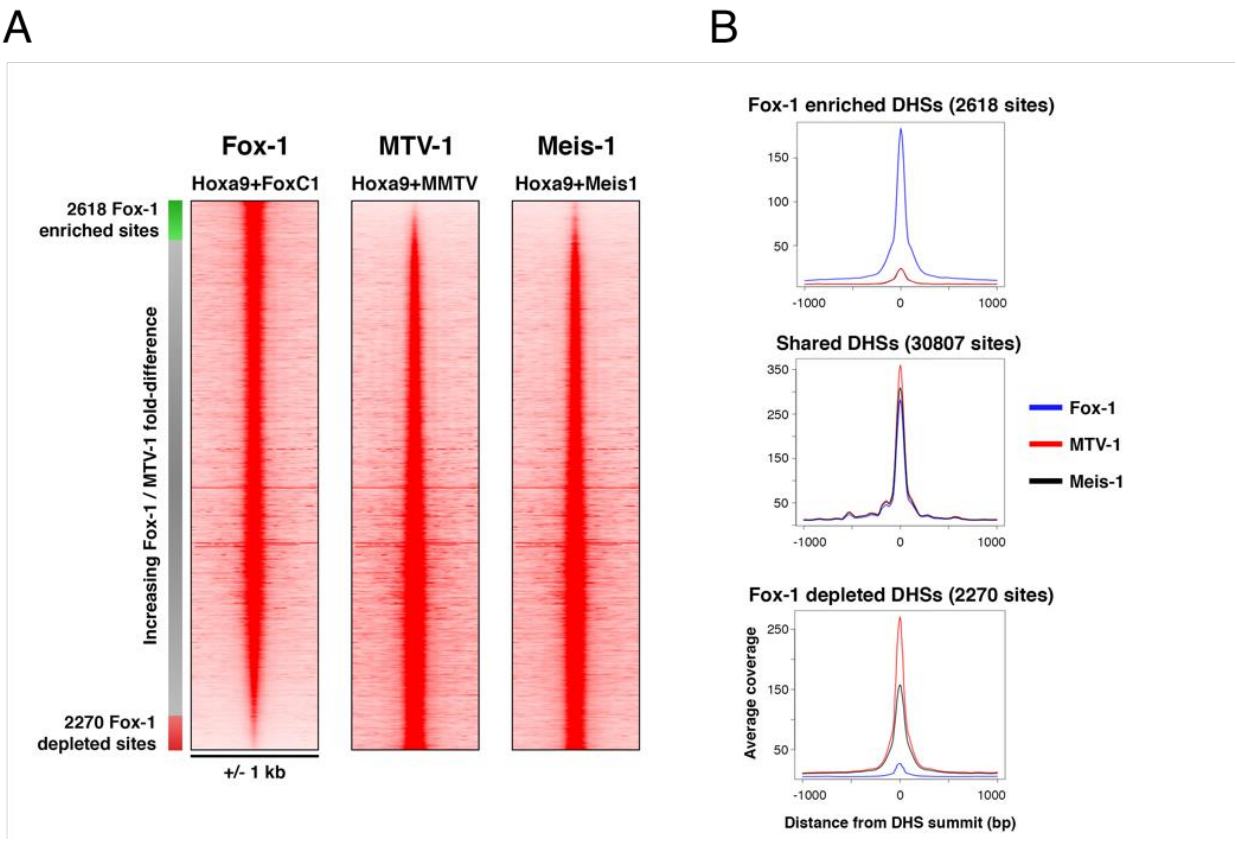


Figure 3:7 **Hoxa9 and FoxC1 activate a specific chromatin signature in mouse AMLs.**

A: Profiles of DNase-seq signal for the anchor samples for each pairwise oncogene expression condition (Fox-1, MTV-1, Meis-1), presented as 2 kb windows centred on the summits of distal DHSs. This analysis includes the union of all peaks in either Hoxa9+FoxC1 or Hoxa9+MMTV, ranked in order of increasing Fox-1 / MTV-1 DNase-seq signal fold-difference. DHSs exhibiting greater than a +/- 4-

fold change were classified as differential DHSs. Populations of differential DHSs enriched and depleted in Hoxa9+FoxC1 AMLs are indicated to the left.

B: Average DNase-seq peak coverage profiles for the three populations of DHSs identified in **A**.

Comparisons between these anchor DNase-seq samples revealed a substantial degree of FoxC1-specific changes in chromatin accessibility. Firstly, co-expression of Hoxa9 with FoxC1 was associated with the specific 4-fold upregulation of 2618 distal DHSs and 4-fold downregulation of 2270 DHSs compared to the Hoxa9+MMTV cells (Figure 3:7). Furthermore, the global DHS profile of cells co-transduced with Hoxa9+Meis1 looked remarkably similar to those for Hoxa9+MMTV.

Correlation clustering analysis of DNase-seq data from each of the mouse AML samples confirmed that FoxC1-dependent differences in global chromatin accessibility was a recurring feature seen in all three biological replicates for each AML sample (Figure 3:8). These findings were underlined by the observation that the global DHS profiles for biological replicates of Hoxa9+MMTV and Hoxa9+Meis1 samples closely correlated with each other in clustering analyses, but not with the Hoxa9+FoxC1 samples, which instead formed a single discrete cluster (Figure 3:8). Finally, the FoxC1-specific accessible chromatin signature presented in Figure 3:7 using the anchor samples Fox-1, MTV-1 and Meis-1 was confirmed to be representative of other biological replicates studied (Figure 3:9).

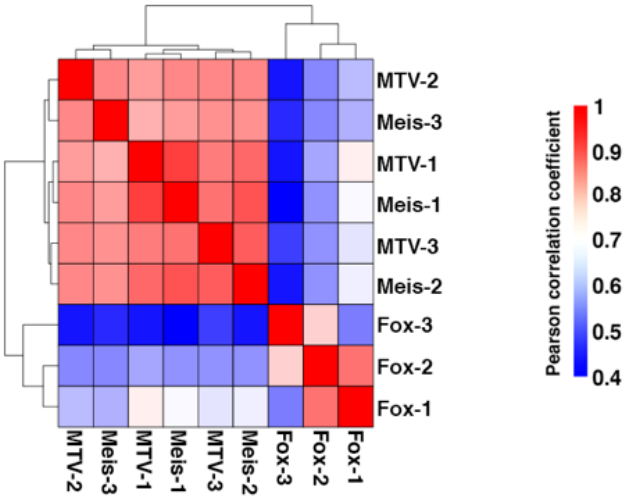


Figure 3:8 **Hoxa9+FoxC1 AMLs feature a common accessible chromatin pattern that is distinct from both Hoxa9+Meis1 and Hoxa9+MMTV cells.**

Hierarchical clustering of the Pearson correlation coefficient scores for distal DHSs across several biological replicates reveals two discrete clusters.

The first comprises the Hoxa9+FoxC1 AML samples, which are less similar to either Hoxa9+Meis1/Hoxa9+MMTV samples, which instead form a separate cluster together.

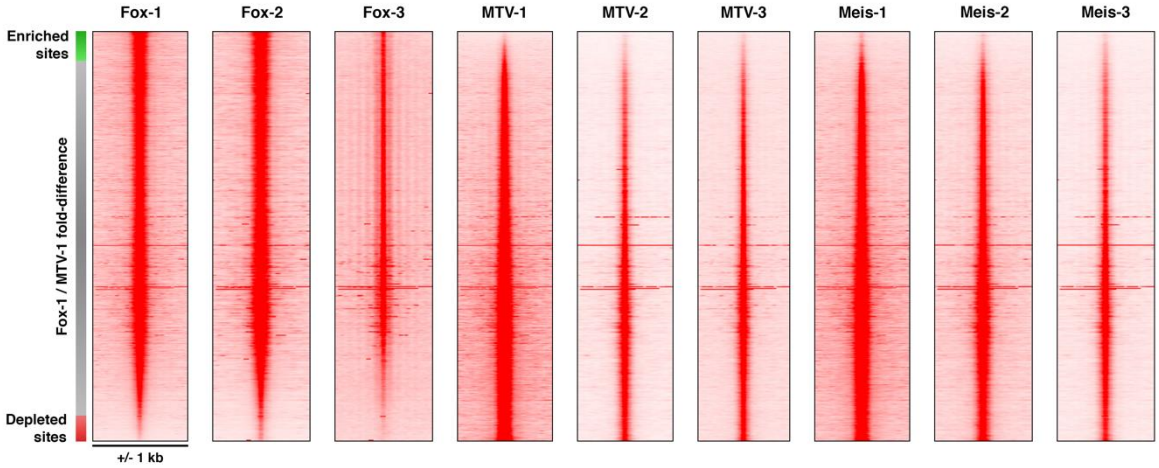


Figure 3:9 **The global DNase-seq distal DHS patterns of samples is consistent across multiple biological replicates.**

DNase-seq profiles are plotted as 2 kb windows, centred on the distal DHS summits. DHSs in all samples are ranked against the same genomic co-ordinates, sorted by increasing Fox-1 / MTV-1 DNase-seq signal fold difference, as in figure 3:6A.

3.4 FoxC1-specific DHSs are enriched for distinct classes of TF binding motifs

To investigate whether particular TFs might contribute to the establishment of the chromatin patterns defined in the previous section, we performed *de novo* motif discovery on the FoxC1-specific populations of enriched and depleted distal DHSs using HOMER[314].

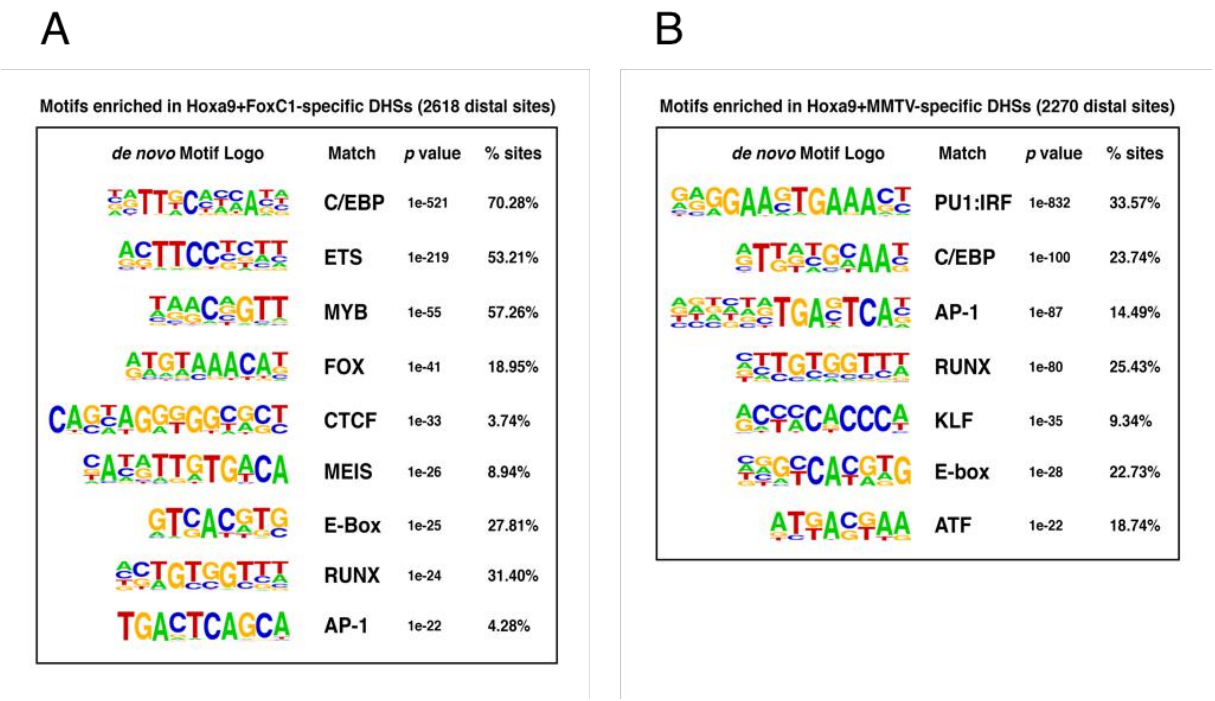


Figure 3:10 **Hoxa9+FoxC1 AMLs are enriched for distinct classes of TF motifs as compared to Hoxa9+MMTV samples.**

A: Results of a HOMER *de novo* motif discovery analysis performed on 2,618 Hoxa9+FoxC1-specific distal DHSs.

B: Results of a similar analysis as (A), but performed on 2,270 distal DHSs specific to Hoxa9+MMTV samples.

First, focussing on the 2,618 distal DHSs enriched in Hoxa9+FoxC1 AML cells, we observed a differential enrichment of FOX binding motifs, consistent with ectopic expression of *FOXC1* in these samples as compared to the other two mouse models (Figure 3:10). Furthermore, a substantial proportion of Hoxa9+FoxC1-specific DHSs

were also distinguished by binding sites for MYB and MEIS family proteins, both of which are expressed in normal myeloid precursors[335]. Finally, we also found an overrepresentation of binding motifs for other myeloid-associated TFs, including RUNX, C/EBP and ETS family proteins.

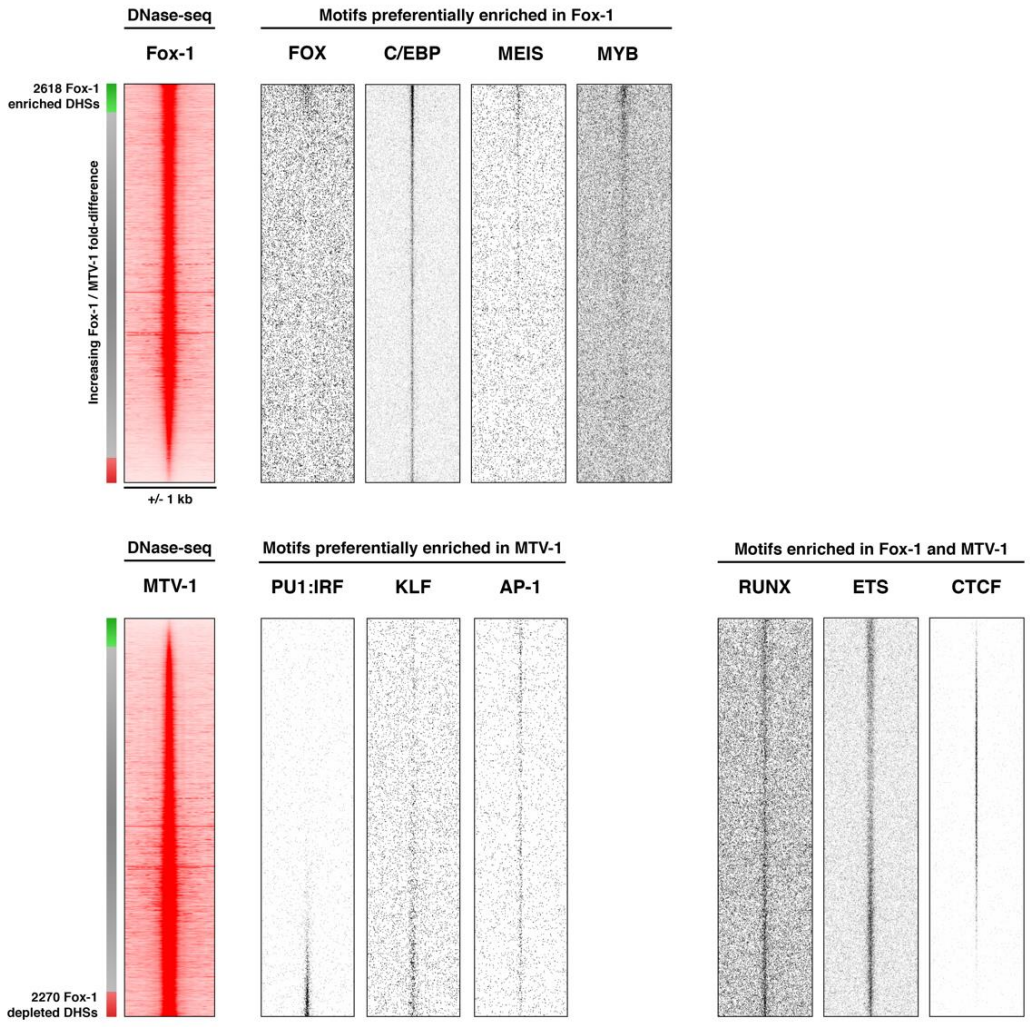


Figure 3:11 **Hoxa9+FoxC1-specific DHSs are enriched for specific TF binding motifs**

Alignment of HOMER *de novo* motif co-ordinates to DNase-seq data sorted by increasing Fox-1/MTV-1 fold-difference reveals enrichment of TF binding motifs which are Hoxa9+FoxC1-specific (top), Hoxa9+MMTV-specific (bottom left) and commonly enriched in both samples (bottom right).

Subsequent analyses looking at the global distribution patterns of these motifs revealed that DHSs containing FOX, MYB, MEIS and C/EBP motifs were most selectively

enriched in the Hoxa9+FoxC1 AMLs and not Hoxa9+MMTV cells (Figure 3:11). Conversely, PU.1:IRF motifs were selectively enriched in Hoxa9+MMTV cells, while RUNX and ETS sites were distributed more uniformly across all three groups of DHSs.

Analyses of the 2,270 distal DHSs specific to Hoxa9+MMTV revealed differential enrichment for binding motifs of TFs classically involved in myeloid cell function or differentiation, including AP-1 and KLF family TFs, and IRF:PU-1 composite sites[161,336,337] (Figure 3:10). Similar to the Hoxa9+FoxC1-upregulated DHSs, Hoxa9+MMTV-specific DHSs also featured substantial enrichment of binding sites for ETS and C/EBP family proteins, although the enrichment of C/EBP motifs was much less substantial than in Hoxa9+FoxC1-specific sites.

Finally, the 30,807 DHSs identified as common to Hoxa9+FoxC1 and Hoxa9+MMTV cells were enriched for CTCF sites, in addition to the RUNX and ETS sites that were detectable in all three groups of DHSs.

3.5 DHSs upregulated in Hoxa9+FoxC1 AML cells are defined by co-localisation of specific TF binding sites

To examine trends in motif enrichment revealed by our *de novo* discovery analyses, we performed motif bootstrapping to reveal significant co-localisation of different TF binding sites within the group of 2,618 Hoxa9+FoxC1-specific DHSs (Figure 3:10). These analyses identified a cluster of FOX, HOX, C/EBP, FOX:E and MYB motifs which significantly co-localised in DHSs specific to Hoxa9+FoxC1 cells, but not DHSs shared with or specific to either Hoxa9+FoxC1 or Hoxa9+MMTV cells. We performed a similar parallel analyses of the 2,270 Hoxa9+MMTV-specific DHSs. The enrichment of PU.1:IRF binding motifs within these DHSs was so strong that any co-localisation patterns were initially difficult to detect (Appendix 3:1). After excluding this motif from a subsequent

analysis, we determined that Hoxa9+MMTV-specific DHSs were also distinguished by co-localisation of HOX motifs with RUNX and AP-1 sites (Figure 3:12).

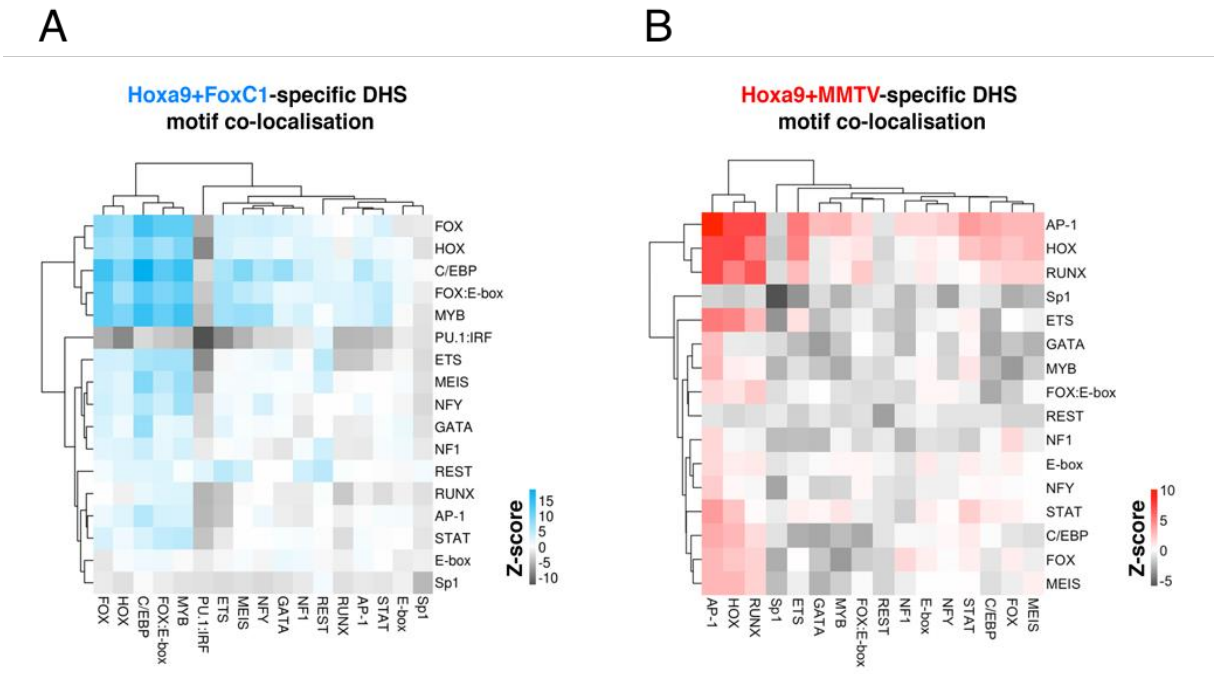


Figure 3:12 **Motif bootstrapping analysis reveals specific patterns of TF binding site co-localisation in Hoxa9+FoxC1 AML-specific DHSs.**

A: Hoxa9+FoxC1 AML cells are defined by a cluster of co-localised FOX, HOX, C/EBP, FOX:E and MYB motifs within the group of 2618 Hoxa9+FoxC1-upregulated DHSs described in Figure 3:7.

B: Hoxa9+MMTV cells instead are distinguished by a cluster of HOX, AP-1 and RUNX motifs within the specific group of 2270 Hoxa9+MMTV-upregulated DHSs described in Figure 3:7.

3.6 Hoxa9+FoxC1 AML cells feature distinctive transcription factor occupancy patterns

To identify specific changes in TF occupancy states between the different mouse models, we re-sequenced the DNase-seq libraries for the three anchor samples (MTV-1, Meis-1 and Fox-1) at high read-depth to permit digital DNase footprinting[338]. These analyses identified a total of 109,421 TF footprints, of which 26,292 were uniquely present in Hoxa9+FoxC1 cells, while 37,416 footprints were commonly present in all three samples (Figure 3:13).

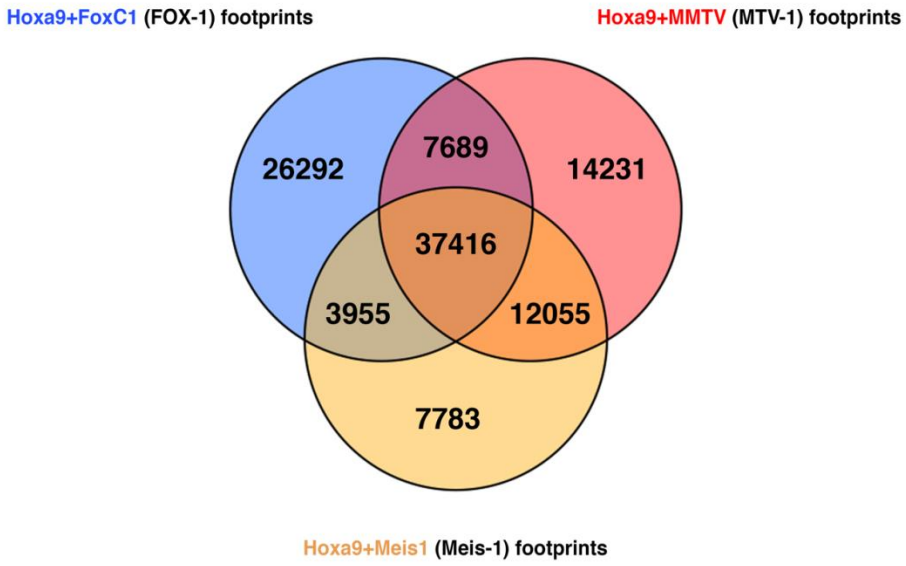


Figure 3:13 **Analyses of digital footprinting data reveals specific groups of differentially footprinted sites in Hoxa9+FoxC1, Hoxa9+Meis1 and Hoxa9+MMTV samples**

Venn diagram showing the overlap of footprints identified in Hoxa9+FoxC1, Hoxa9+Meis1 and Hoxa9+MMTV cells.

De novo motif discovery analyses performed on the Hoxa9+FoxC1-specific footprints revealed an overwhelming preferential abundance of occupied C/EBP (26%) and CTCF (11%) sites, together with more modest numbers of occupied Sp1 and ETS sites (~3% of sites). In addition, we observed a very low frequency of occupied REST and NF1 motifs (less than 1%) in the Hoxa9+FoxC1-specific footprints that were not found in either the Hoxa9+MMTV or Hoxa9+Meis1-specific groups (Figure 3:14). In contrast, Hoxa9+MMTV-specific and Hoxa9+Meis1-specific footprints were both preferentially enriched for AP-1, KLF and E-box sites instead. Surprisingly, the Hoxa9+FoxC1 footprints did not feature a significant enrichment of FOX sites. Furthermore, the footprinting analysis identified RUNX motifs in just 0.3% of the Hoxa9+FoxC1-specific footprints, even though 31% of the specific DHSs contained a RUNX motif. This suggests that RUNX1 binding is not easily detected via footprinting.

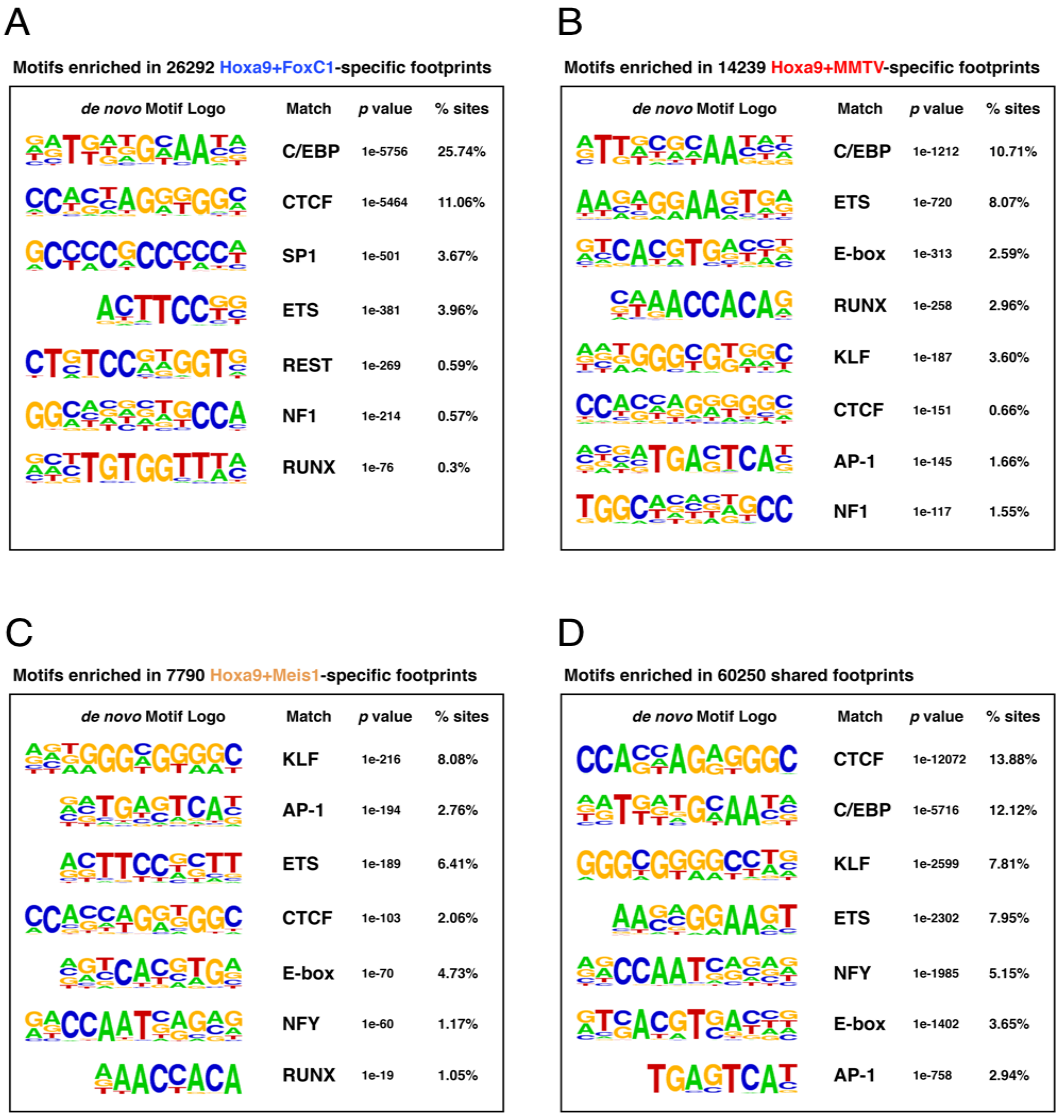


Figure 3:14 **Analyses of digital footprinting data reveals Hoxa9+FoxC1 AML-specific patterns of motif occupancy.**

A-C: HOMER *de novo* motif result for footprints specific to: Hoxa9+FoxC1 (**A**); Hoxa9+MMTV (**B**); Hoxa9+Meis1 (**C**), or; footprints present in at least two of Hoxa9+FoxC1, Hoxa9+MMTV or Hoxa9+Meis1 (**D**).

3.7 FoxC1 expression is associated with global changes in gene expression

Given that forced expression of FoxC1 with Hoxa9 led to extensive specific changes in the accessible chromatin features of transduced cells, we sought to investigate the impact on gene expression arising from the activity of these proto-oncogenes by performing RNA-seq analyses on a subset of the same AML samples. Consistent with comparative DNase-seq analysis, Hoxa9+FoxC1 AML samples featured a distinct gene expression profile, forming a discrete cluster in correlation clustering analyses with the other samples (Figure 3:15). These findings were corroborated by a principal component analysis of individual RNA-seq samples, which demonstrated that Hoxa9+FoxC1 samples formed a closely related group that was separated from Hoxa9+Meis1 and Hoxa9+MMTV samples along the first principal component (Figure 3:15).

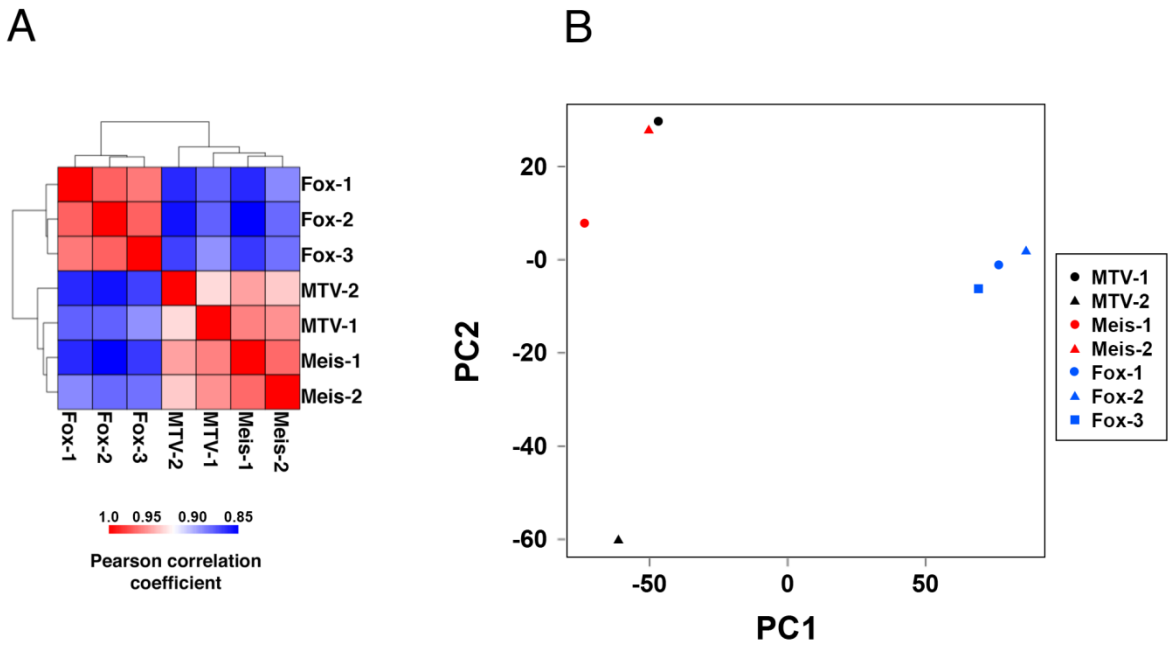


Figure 3:15 **Hoxa9+FoxC1 AMLs feature a specific gene expression pattern.**

A: Hierarchical clustering of Pearson correlation coefficient scores of RNA-seq data generated for each pairwise oncogene expression condition.

B: Principal component analysis of the same data as in (A).

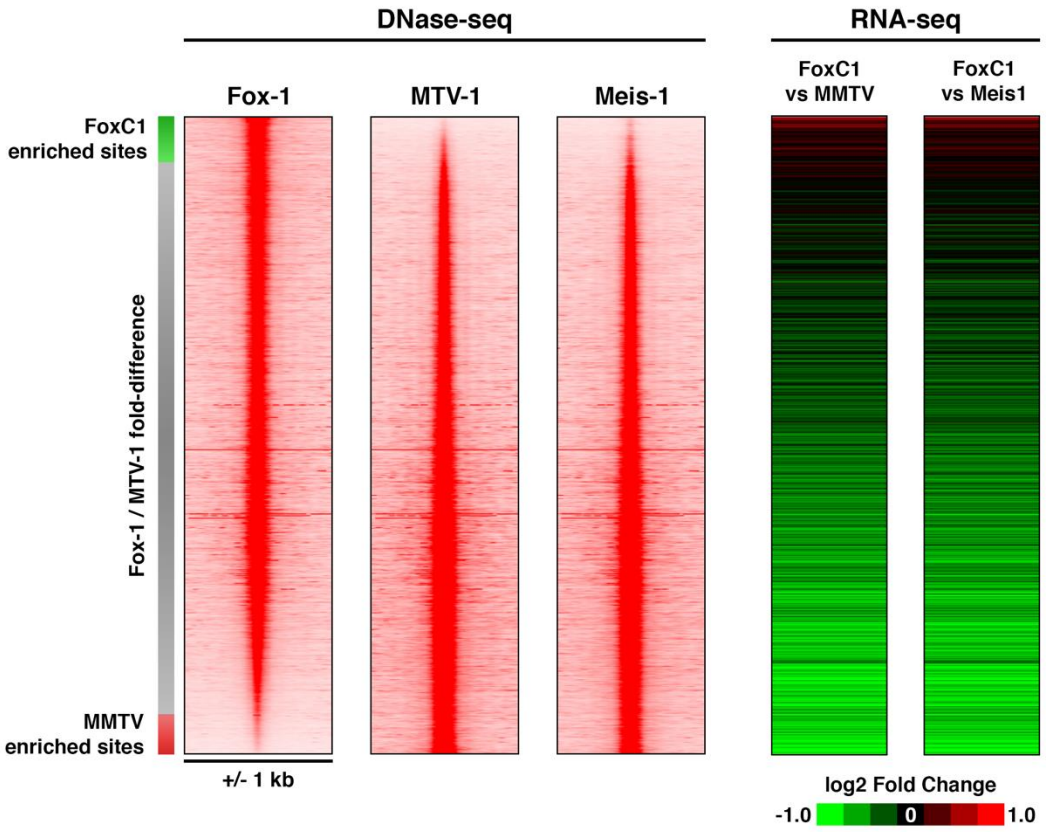


Figure 3:16 **FoxC1-specific changes in chromatin accessibility are associated with a specific pattern of differential gene expression.**

To the left are DNase-seq profiles ranked by increasing Fox-1/MTV-1 DNase-seq signal fold difference, shown in 2 kb windows centred on the peak of the DHS summits. To the right are the corresponding changes in average RNA expression for the nearest genes of the DHS summits. Changes in RNA expression are calculated as fold changes of Hoxa9+FoxC1 AMLs relative to either Hoxa9+MMTV or Hoxa9+Meis1 samples.

3.8 FoxC1-dependent changes in chromatin accessibility correlate with specific changes in gene expression

Next, we sought to examine how changes in chromatin accessibility specific to Hoxa9+FoxC1 AMLs related to observed shifts in gene expression by integration of our DNase-seq and RNA-seq data (Figure 3:16). These analyses revealed widespread FoxC1-dependent downregulation of genes associated with nearby DHSs that were either depleted in Hoxa9+FoxC1 AMLs or present in all three mouse models.

Conversely, the most Hoxa9+FoxC1-enriched group of DHSs were associated with the specific upregulation of nearby genes. This pattern was strikingly consistent, irrespective of whether gene expression patterns in the Hoxa9+FoxC1 samples were compared to Hoxa9+Meis1 or Hoxa9+MMTV samples. This observation was underlined further by Venn diagram analyses demonstrating substantial overlaps of differentially expressed genes in Hoxa9+FoxC1 AMLs as compared to either Hoxa9+Meis1 or Hoxa9+MMTV samples (Figure 3:17).

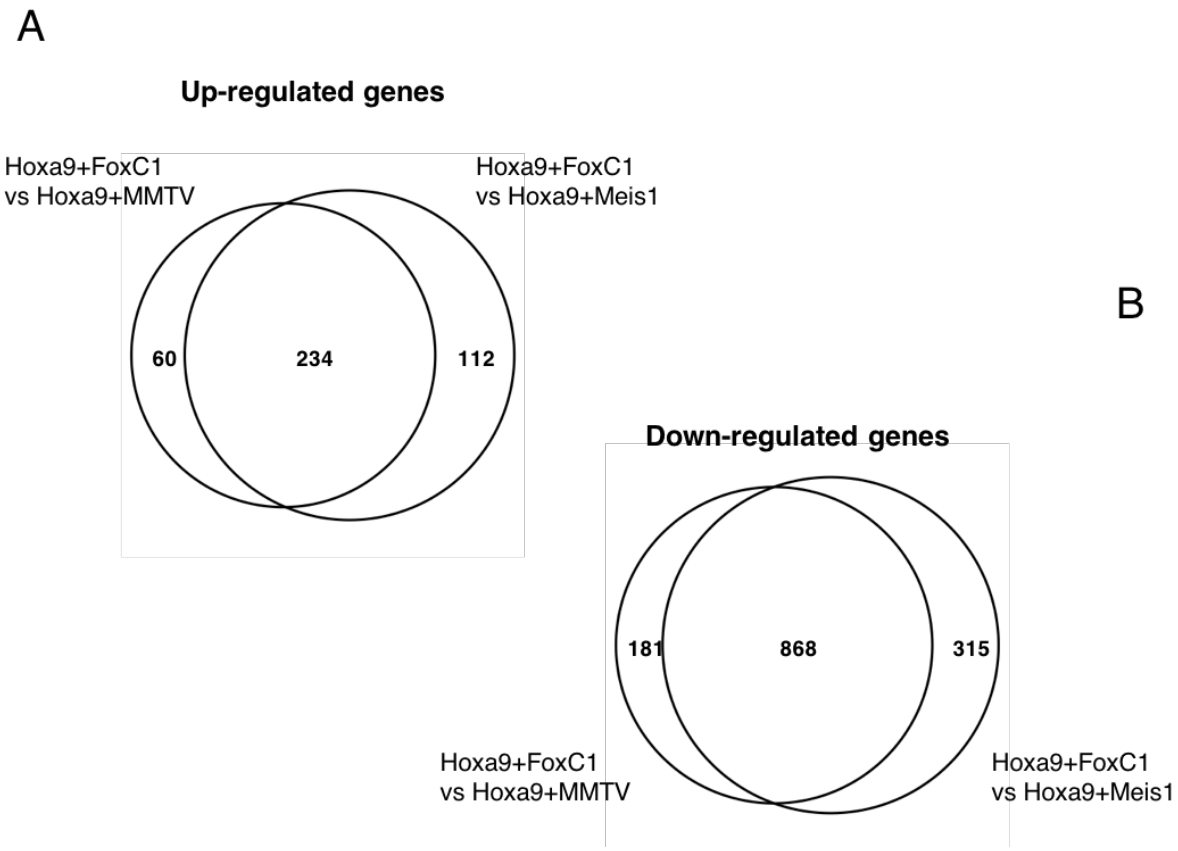


Figure 3:17 **FoxC1-dependent changes in gene expression are broadly similar when compared to either Hoxa9 alone or Hoxa9+Meis1 AML cells.**

A: Overlap of up-regulated genes from analysed RNA-seq data in Hoxa9+FoxC1 samples relative to either Hoxa9+MMTV or Hoxa9+Meis1.

B: Overlap of down-regulated genes between Hoxa9+FoxC1 samples and either Hoxa9+MMTV or Hoxa9+Meis1.

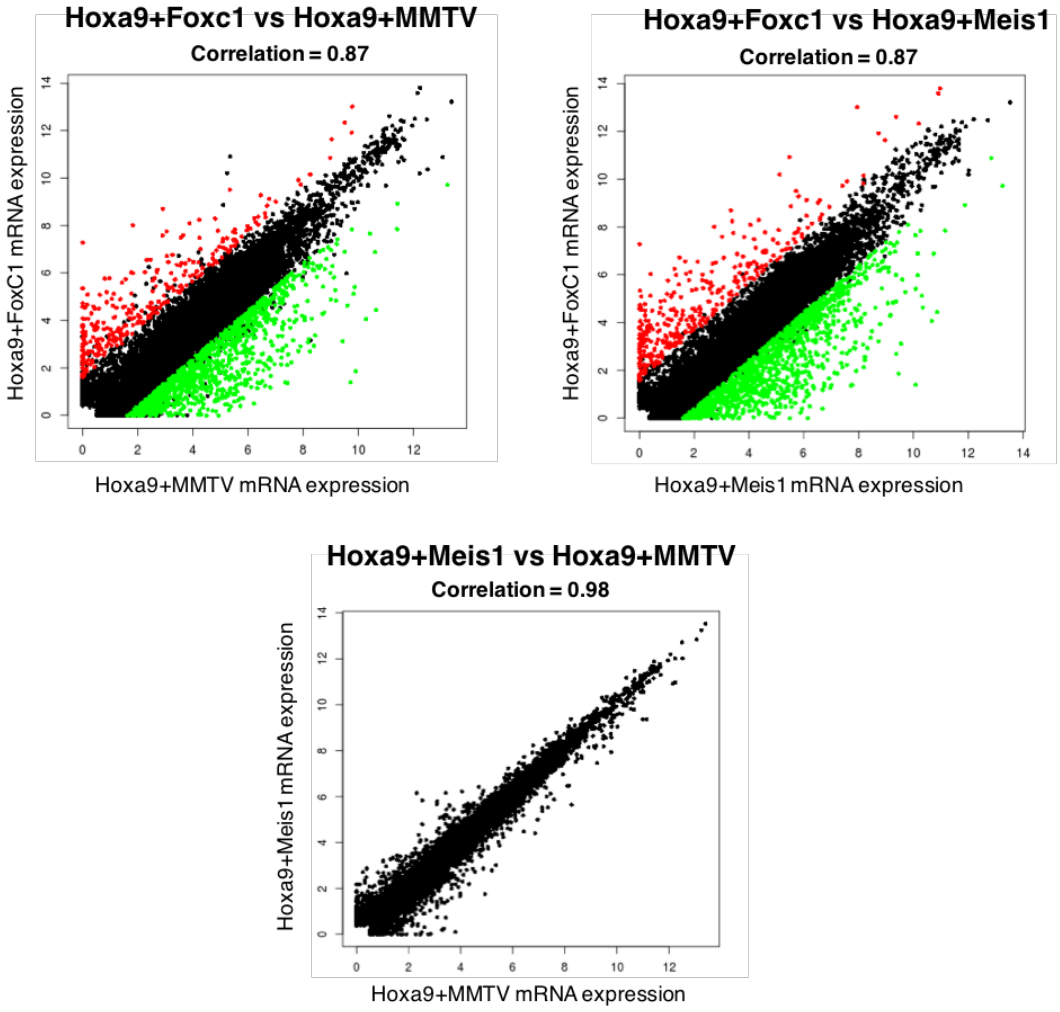


Figure 3:18 **Pairwise comparisons of RNA-seq data reveal FoxC1-dependent changes in gene expression.**

XY scatter plots of normalised mRNA expression values [log2(TPM+1)] from RNA-seq data are shown as pairwise comparisons between the three ectopic oncogene expression combinations. Significantly upregulated genes are plotted in red, and significantly downregulated genes in green. Genes not significantly changing in expression between two groups are plotted in black.

To confirm trends in mRNA expression patterns revealed by the above global analyses, we performed pairwise comparisons of RNA-seq data to identify differentially expressed genes between two biological replicates of all three mouse models. These analyses revealed a comparatively weaker correlation between Hoxa9+FoxC1 AMLs and either Hoxa9+Meis1 or Hoxa9+MMTV (Figure 3:18). Furthermore, Hoxa9+Meis1 AMLs and

Hoxa9+MMTV samples correlated more robustly with each other. In fact, when taking adjusted *p* values (<0.05) into account, no genes were differentially expressed at a significant level between Hoxa9+Meis1 AMLs and Hoxa9+MMTV cells.

3.9 FoxC1 and Hoxa9 repress the expression of genes associated with normal myeloid cell differentiation and innate immunity

To gain an insight into pathways significantly deregulated by FoxC1 and associated target genes in mouse AML, we performed KEGG pathway analyses[339] on sets of significantly differentially up- and downregulated genes in Hoxa9+FoxC1 AMLs. Strikingly, very few KEGG pathway terms were significantly enriched in the upregulated gene list for Hoxa9+FoxC1 AMLs as compared to either Hoxa9+Meis1 or Hoxa+MMTV cells (Table 3:1). With the exception of “metabolic pathways”, none of the terms annotated to Hoxa9+FoxC1-upregulated genes seemed to be of direct relevance to leukaemia.

By contrast, KEGG pathway analyses of Hoxa9+FoxC1-downregulated genes as compared to Hoxa9+MMTV and Hoxa9+Meis1 revealed several statistically-significant (adj. *p* <0.05) KEGG annotations overtly related to the AML pathogenesis. For example, terms normally associated with terminally differentiated myeloid cells, including “Toll-like receptor signalling", “Cytokine-cytokine receptor interaction” and “NF-κB signalling" were all downregulated in Hoxa9+FoxC1 AMLs, regardless of whether the differential gene expression comparison was made against Hoxa9+Meis1 AMLs or Hoxa9+MMTV. Furthermore, manual inspection of genes annotated with these terms confirmed comprehensive downregulation of several critical mediators of normal monocyte and macrophage function, including the C-C motif chemokine ligand genes *Ccl3* and *Ccl4*, and cell-surface receptors associated with monocyte/macrophage differentiation, *Cd14* and *Csf1r*. Together with a marked decrease in the expression of mature myeloid

markers, we observed downregulation of 5 IRF-family and 3 KLF-family TF genes, which promote terminal differentiation of myeloid and/or lymphoid lineages in normal haematopoiesis[340,341] (Figure 3:19). This was accompanied by a modest downregulation of the tumour suppressor gene *Foxo1*, which was previously documented as a direct activator of *Klf4* during lymphoid development[342,343]. Finally, we observed a modest but significant downregulation of the tumour suppressor genes *Rb1* and *Pten* in Hoxa9+FoxC1 AMLs as compared to either Hoxa9+MMTV or Hoxa9+Meis1.

Table 3:1 **KEGG pathway analyses of differentially regulated genes in Hoxa9+FoxC1 AMLs.**

KEGG pathway enrichment analyses were performed on all significantly up- or downregulated genes in Hoxa9+FoxC1 RNA-seq datasets, as compared to either Hoxa9+MMTV or Hoxa9+Meis1.

Top KEGG pathways significantly up-regulated in Hoxa9+FoxC1 (vs Hoxa9+MMTV)

ID	Term	Count	*p* value	Adj. *p* value
mmu01100	Metabolic pathways	59	1.36e-05	0.0031

Top KEGG pathways significantly up-regulated in Hoxa9+FoxC1 (vs Hoxa9+MMTV)

ID	Term	Count	*p* value	Adj. *p* value
mmu05322	Systemic lupus erythematosus	21	8.13e-08	1.92E-05
mmu01100	Metabolic pathways	70	2.00e-05	0.0023
mmu05034	Alcoholism	19	1.42e-04	0.0110

Top 10 KEGG pathways significantly down-regulated in Hoxa9+FoxC1 (vs Hoxa9+MMTV)

ID	Term	Count	*p* value	Adj. *p* value
mmu04620	Toll-like receptor signalling pathway	26	1.03e-10	2.46e-08
mmu04380	Osteoclast differentiation	23	1.10e-06	1.31e-04
mmu05202	Transcriptional misregulation in cancer	26	3.50e-06	2.78e-04
mmu04668	TNF signalling pathway	19	2.33e-05	0.0014
mmu04621	NOD-like receptor signalling pathway	13	3.96e-05	0.0019
mmu04064	NF κB signalling pathway	17	6.70e-05	0.0027
mmu04672	Intestinal immune network for IgA production	11	6.75e-05	0.0023
mmu04060	Cytokine-cytokine receptor interaction	30	6.89e-05	0.0020
mmu04640	Hematopoietic cell lineage	15	1.69e-04	0.0045
mmu05152	Tuberculosis	23	2.38e-04	0.0057

Top 10 KEGG pathways significantly down-regulated in Hoxa9+FoxC1 (vs Hoxa9+Meis1)

ID	Term	Count	*p* value	Adj. *p* value
mmu04620	Toll-like receptor signalling pathway	29	2.44e-11	5.91e-09
mmu04640	Hematopoietic cell lineage	22	5.83e-08	7.06e-06
mmu04060	Cytokine-cytokine receptor interaction	39	4.13e-07	3.33e-05
mmu05144	Malaria	15	1.31e-06	7.90e-05
mmu04380	Osteoclast differentiation	25	1.56e-06	7.56e-05
mmu04668	TNF signaling pathway	21	2.07e-05	8.35e-04
mmu04142	Lysosome	22	3.53e-05	0.0012
mmu04621	NOD-like receptor signalling pathway	14	4.71e-05	0.0014
mmu05202	Transcriptional misregulation in cancer	26	6.75e-05	0.0018
mmu04510	Focal adhesion	29	1.73e-04	0.0042

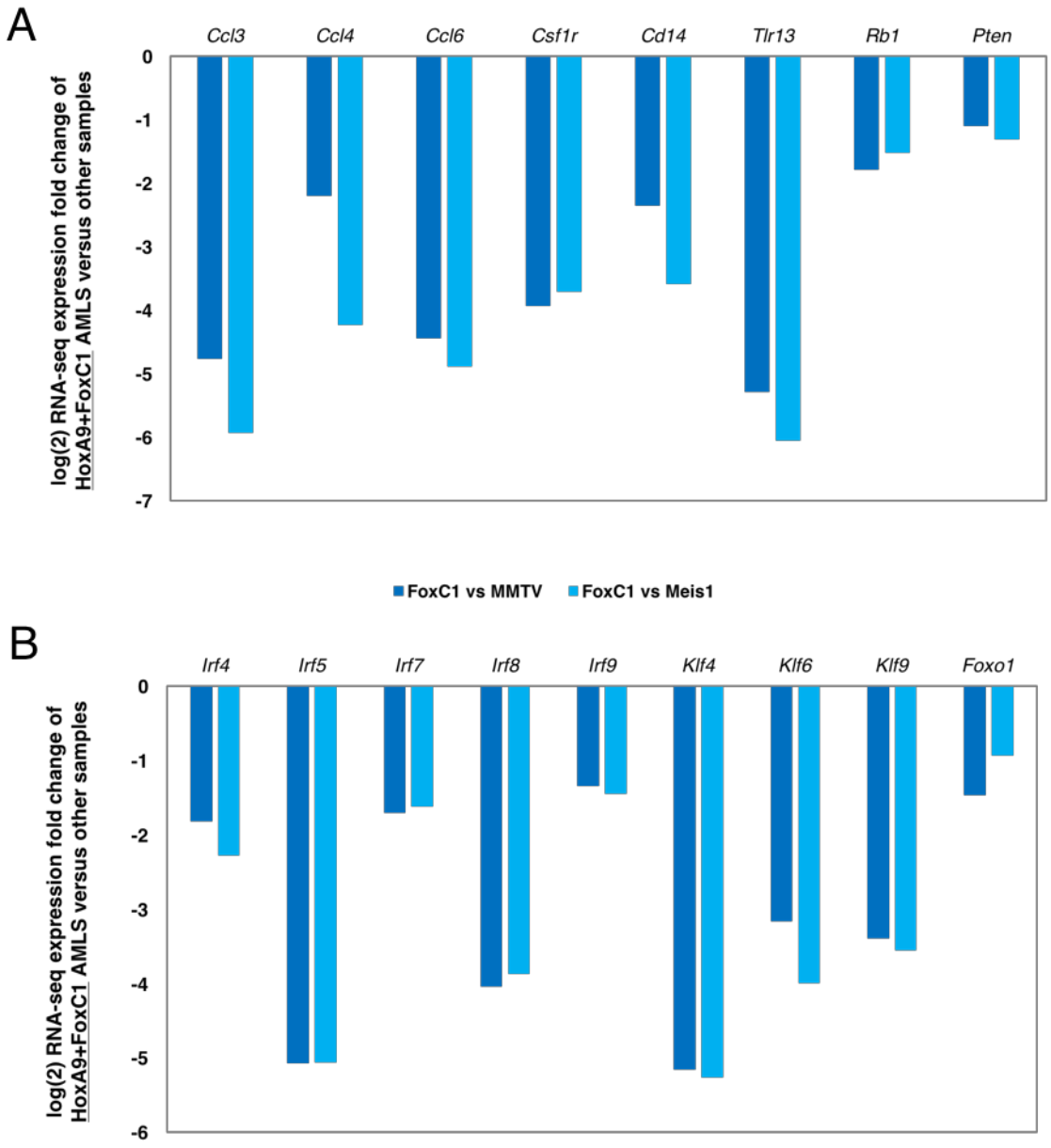


Figure 3:19 **Genes downregulated in Hoxa9+FoxC1 AMLs include myeloid-specific markers, tumour suppressors and transcription factors involved in terminal haematopoietic differentiation.**

A: Downregulation of myeloid-specific marker genes and the tumour suppressors *Rb1* and *Pten* in Hoxa9+FoxC1 AMLs relative to Hoxa9+MMTV or Hoxa9+Meis1.

B: Downregulated TF genes expressed during myeloid or lymphoid differentiation in Hoxa9+FoxC1 AMLs as compared to Hoxa9+MMTV or Hoxa9+Meis1.

mRNA expression changes are calculated from normalised RNA-seq data as the average fold change between Hoxa9+FoxC1 AMLs and either Hoxa9+MMTV (dark blue) or Hoxa9+Meis1 (light blue). All fold changes are statistically significant (adj. $P < 0.05$).

3.10 FoxC1 and Hoxa9 AMLs feature specific upregulation of transcription factor and leukaemia-associated target genes

Because we observed specific enrichment of TF binding motifs in DHSs upregulated in Hoxa9+FoxC1 AMLs (3.4), we analysed RNA-seq data to identify whether any TF genes encoding the TFs binding these motifs were differentially regulated. In keeping with an overrepresentation of C/EBP, MEIS and MYB motifs, we observed striking Hoxa9+FoxC1-dependent upregulation of *Cebpe* and *Meis2* together with modest upregulation of *Myb* as compared to Hoxa9+Meis1 or Hoxa9+MMTV (Figure 3:20). These analyses also revealed substantial upregulation of the TF gene *Sox5,* and of *L3mbtl2*, which encodes a non-canonical Polycomb regulator[344].

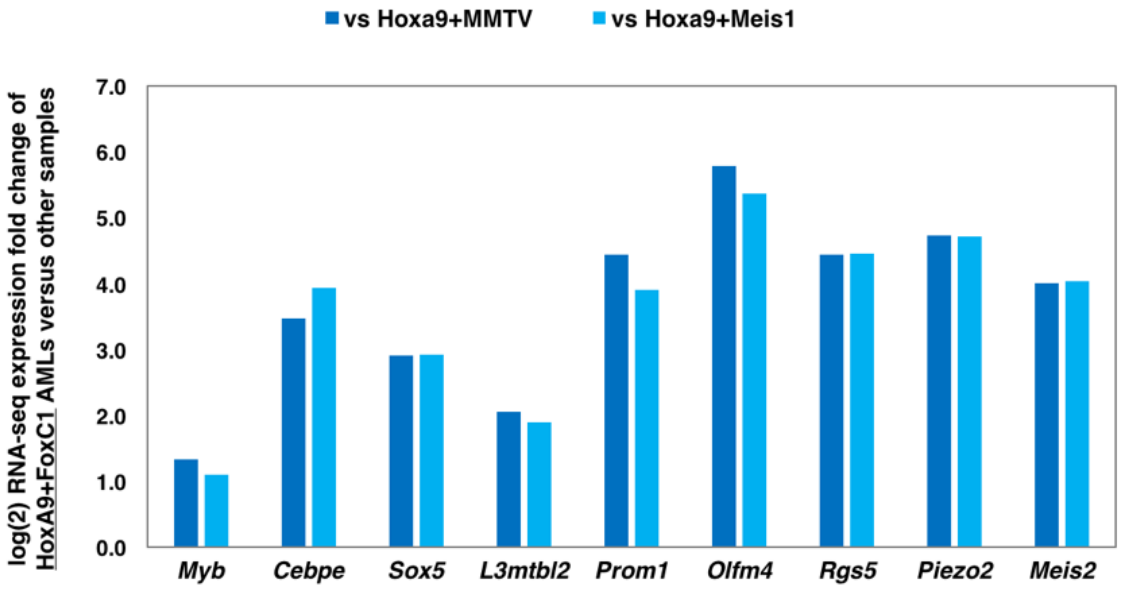
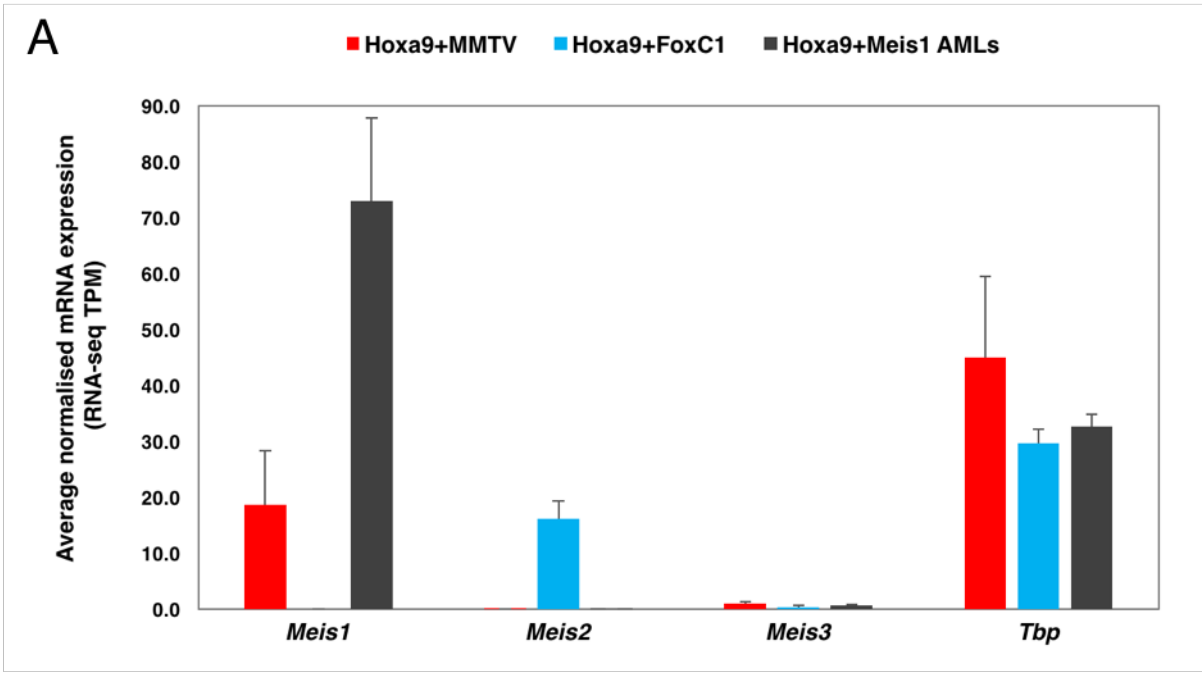


Figure 3:20 **Specific upregulation of mRNA expression in Hoxa9+FoxC1 AMLs**

mRNA expression fold changes for genes upregulated in Hoxa9+FoxC1 AMLs as compared to Hoxa9+Meis1 AMLs or Hoxa9+MMTV, calculated from average normalised RNA-seq expression values across multiple replicates for each condition.



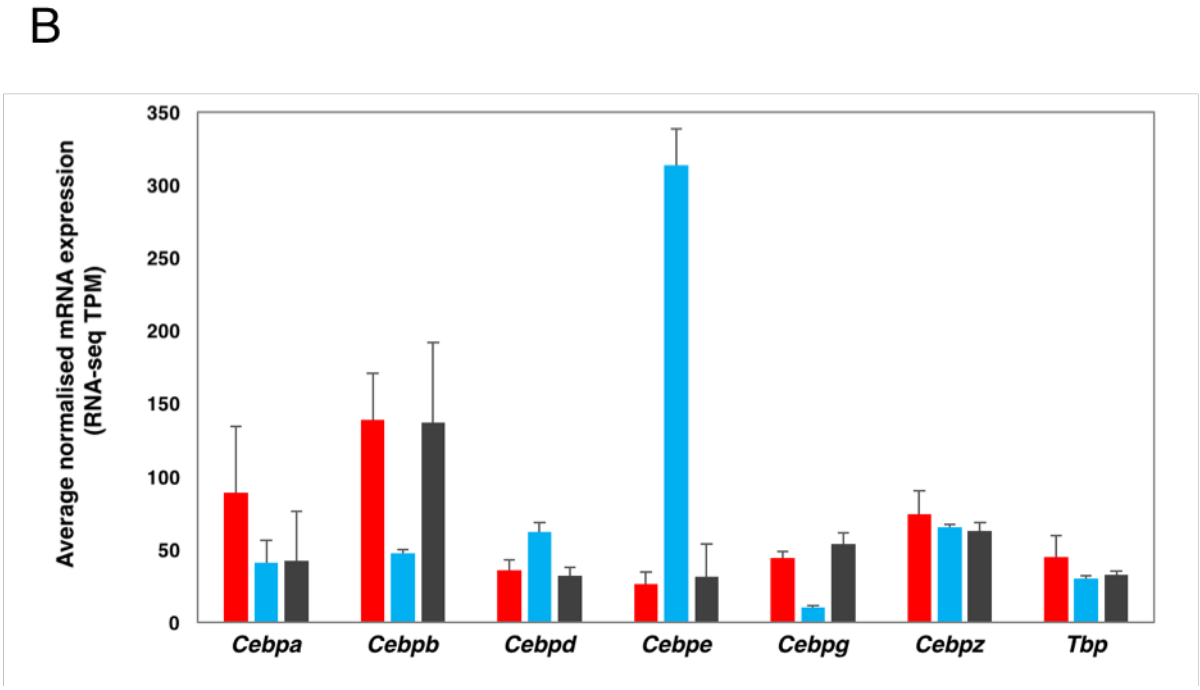


Figure 3:21 ***Meis2* and *Cebpe* are the only substantially upregulated genes of their families in Hoxa9+FoxC1 AMLs.**

A: Analyses of normalised RNA-seq expression values for the three Meis family genes together with a constitutively expressed gene, *Tbp*. Values are presented as the average of three RNA-seq replicates for each of Hoxa9+MMTV, Hoxa9+Meis1 and Hoxa9+FoxC1 AMLs. Error bars were calculated from the standard error of the mean for replicate samples.

B: Similar analysis to A, but for C/EBP family genes, again with *Tbp* as an internal control.

Meis2 and *Cebpe* were among the most highly upregulated genes in Hoxa9+FoxC1 AMLs, both appearing in the top 15% of upregulated genes as compared to Hoxa9+MMTV (log_2 FC ≥ 1.5, adj. $p < 0.05$). Because both of these genes belong to wider families of TFs, we manually inspected the RNA-seq expression levels for each family member (Figure 3:21). These analyses confirmed that both *Cebpe* and *Meis2* were the most substantially upregulated genes in their respective families in Hoxa9+FoxC1 AMLs as compared to either Hoxa9+MMTV or Hoxa9+Meis1 cells. *Cebpd* was the only other C/ebp family gene that showed a modest trend of upregulation in Hoxa9+FoxC1 AMLs, however this was not statistically significant (Hoxa9+FoxC1 vs Hoxa9+MMTV adj. $p = 0.18$). Conversely, the other C/EBP family genes *Cebpb* and *Cebpg* were significantly downregulated in Hoxa9+FoxC1 AMLs (Hoxa9+FoxC1 vs Hoxa9+MMTV adj. $p = 0.008$ and 0.007, respectively), possibly due to feedback regulation. Finally, *Cebpa* showed a trend of downregulation in Hoxa9+FoxC1 AMLs, although this was not statistically significant (Hoxa9+FoxC1 vs Hoxa9+MMTV adj. $p = 0.12$). Similar RNA-seq analyses for the Meis family genes confirmed that only *Meis2* was upregulated in Hoxa9+FoxC1 AMLs as compared to the other samples.

Finally, in addition to Hoxa9+FoxC1-specific upregulation of these TF and chromatin regulator genes, we observed an increase in mRNA expression levels of the LSC marker *Prom1*[345,346], the tumour-specific antigen *Rgs5*[347], the tumour-associated *Olfm4*[348] and *Piezo2*[349], which was previously reported as a diagnostic marker of NPM1-mutant AMLs (Figure 3:20).

3.11 Integrative analyses of DNase-seq and RNA-seq data reveals *cis*-regulatory elements potentially involved in the activation of Hoxa9+FoxC1 target genes

To identify specific DHSs that may be involved in Hoxa9+FoxC1-specific patterns of gene expression, we integrated the overlap between commonly up-/downregulated mRNAs and DHSs that are annotated to the same gene (Figure 3:22).

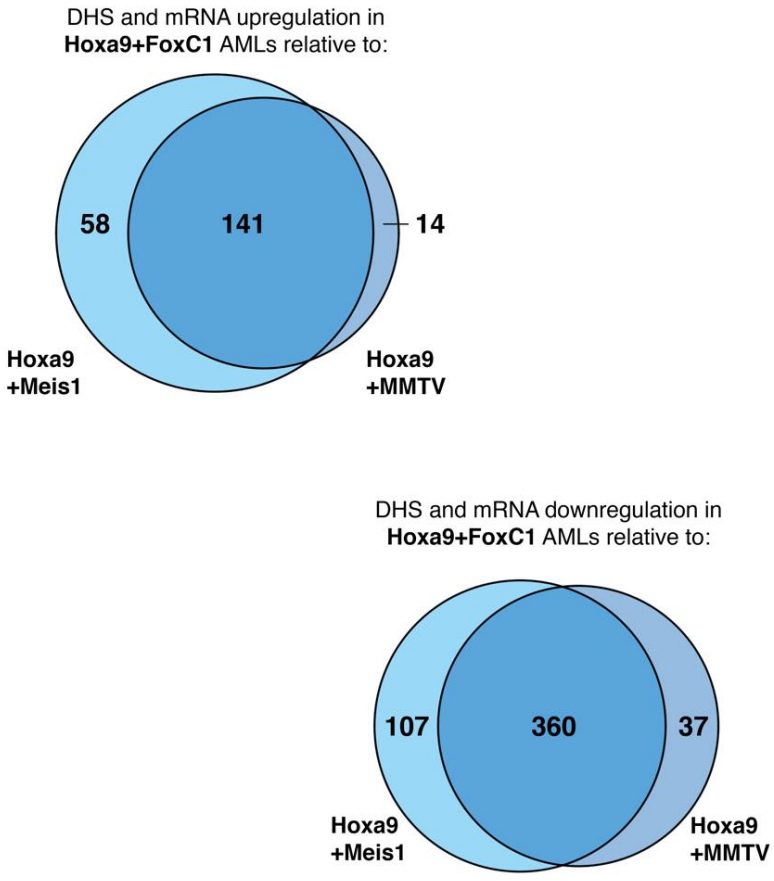


Figure 3:22 **Commonly changing DHSs and gene expression in Hoxa9+FoxC1 AMLs are largely similar whether compared to Hoxa9+Meis1 or Hoxa9 alone.**

Venn diagram analyses showing the overlap of commonly upregulated (top) or downregulated (bottom) DHSs and mRNA expression in Hoxa9+FoxC1 AMLs as compared to either Hoxa9+Meis1 or Hoxa9 alone.

A 4-fold difference in DNase-seq signal was used to define differential DHSs, while a 2-fold difference in RNA-seq signal was used to define differentially expressed genes.

These analyses revealed 141 upregulated genes in Hoxa9+FoxC1 AMLs that also had local upregulated DHSs, compared to both Hoxa9+Meis1 and Hoxa9+MMTV. Conversely, we observed 360 downregulated genes with commonly downregulated local DHSs in Hoxa9+FoxC1 AMLs as compared to both Hoxa9+Meis1 and Hoxa9+MMTV. Furthermore, consistent with previous observations of either DNase-seq or RNA-seq data from these samples, there was substantial overlap in Hoxa9+FoxC1-specific genes, regardless of whether this was relative to Hoxa9+Meis1 AMLs or Hoxa9+MMTV.

Filtering these results for genes only with substantially changing mRNA expression (2 fold up-/downregulated in Hoxa9+FoxC1 vs Hoxa9+MMTV, adj. p <0.05) revealed 57 upregulated genes and 187 downregulated genes in Hoxa9+FoxC1 AMLs (Appendix 3:2).

Manual inspection of several genes identified by these analyses revealed Hoxa9+FoxC1-specific changes in chromatin accessibility that were associated with enhanced gene expression (Figure 3:23 and Figure 3:24). For example, Hoxa9+FoxC1-dependent upregulation of *Myb* expression was associated with enhanced hypersensitivity across the gene locus, as well as the presence of six footprints in Hoxa9+FoxC1 alone. Similar examples were observed at the *Cebpe*, *Prom1*, *Sox5, L3mbtl2* loci and elsewhere, many of which featured evidence of Hoxa9+FoxC1-specific occupancy patterns in the form of digital DNase I footprints.

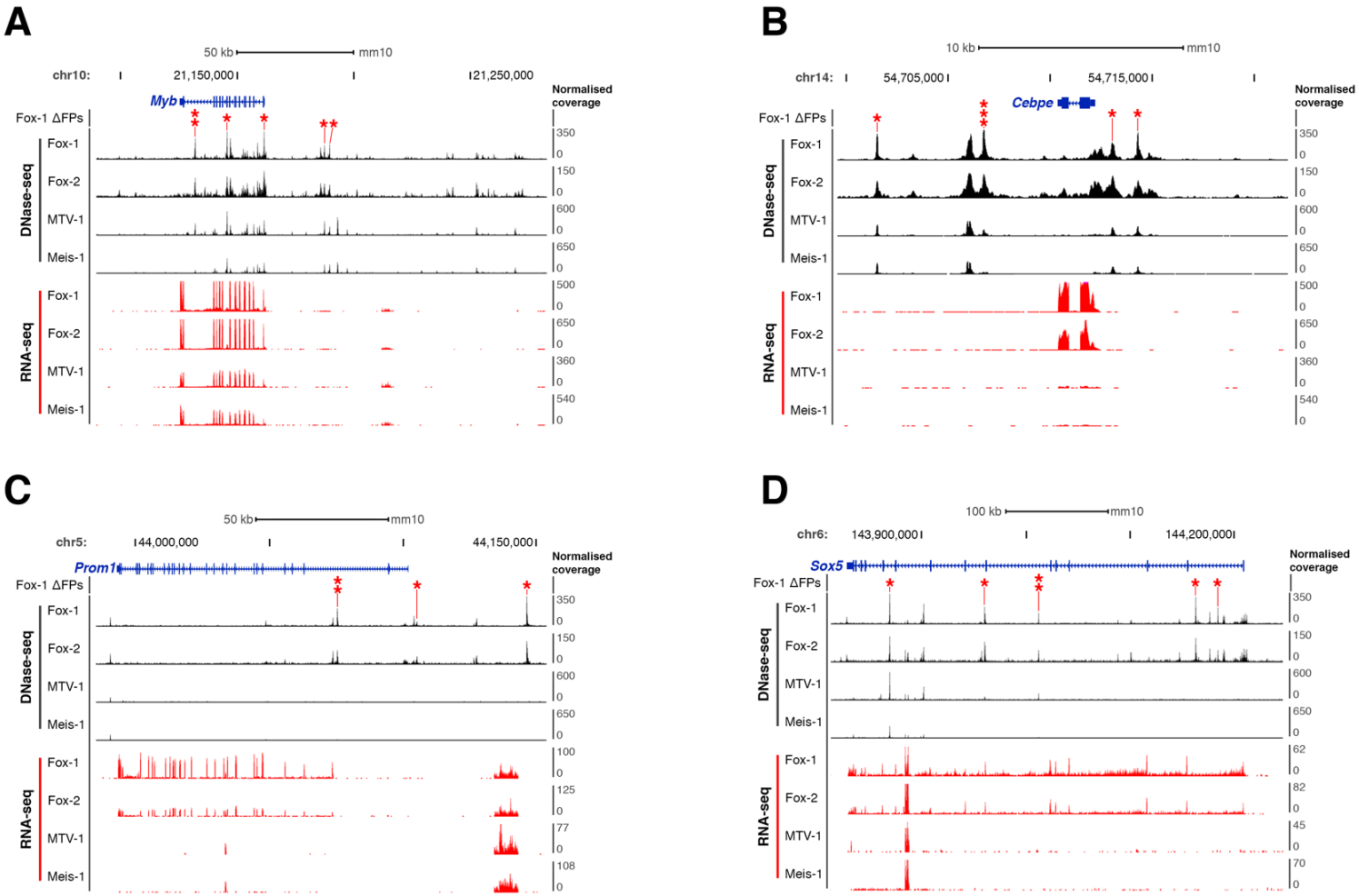


Figure 3:23 **Hoxa9+FoxC1-specific activation of target genes**

A-D: UCSC genome browser views of DNase-seq and RNA-seq data for four genes (*Myb*, *Cebpe*, *Prom1* and *Sox5*) which are specifically upregulated in Hoxa9+FoxC1 AMLs. Specific footprinted sites in Hoxa9+FoxC1 cells (Fox-1 ΔFPs) are indicated above the DHS tracks with red asterixes.

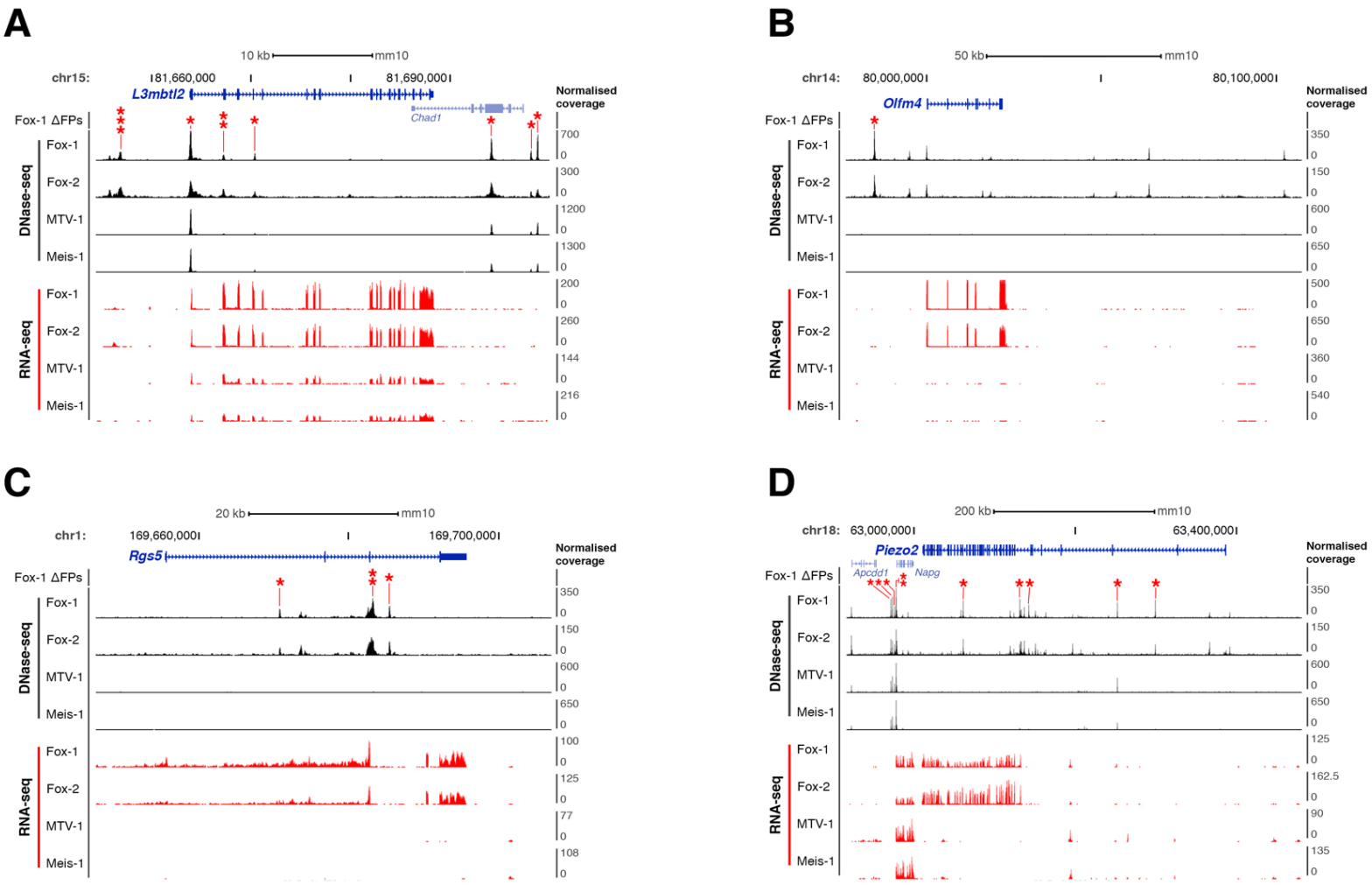


Figure 3:24 **Hoxa9+FoxC1-specific activation of target genes (further examples)**

A-D: UCSC genome browser view of DNase-seq and RNA-seq data for four genes (*L3mbtl2*, *Olfm4*, *Rgs5* and *Piezo2*) which are specifically upregulated in Hoxa9+FoxC1 AMLs. Specific footprinted sites in Hoxa9+FoxC1 cells (Fox-1 ΔFPs) are indicated above the DHS tracks with red asterixes.

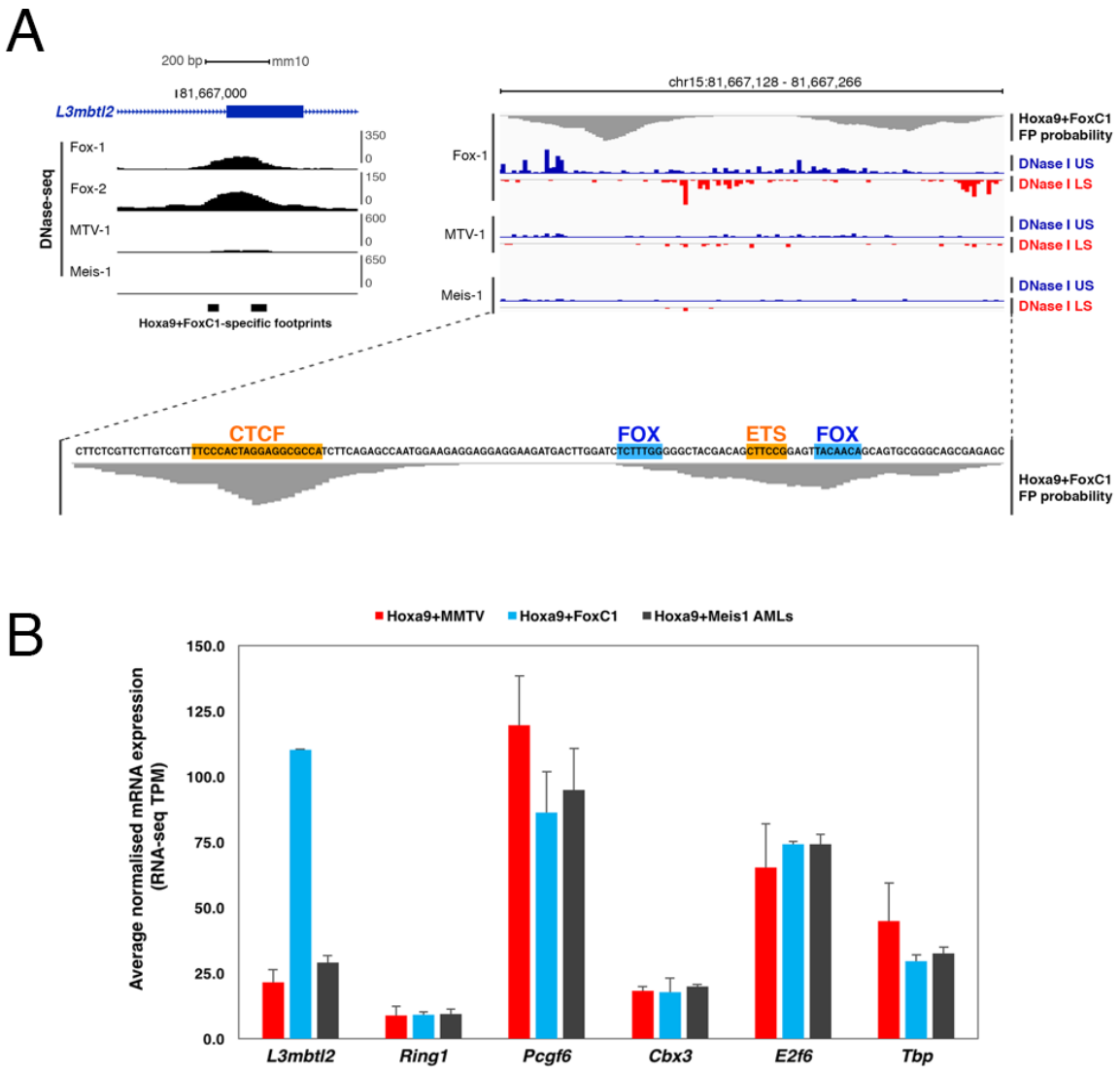


Figure 3:25 **FoxC1 expression is associated with activation of the Polycomb group gene *L3mtbl2*.**

A: (Left) UCSC genome browser snapshot showing a Hoxa9+FoxC1-specific DHS close to the second exon of *L3mbtl2*.

(Right) Detail of Hoxa9+FoxC1-specific patterns of genomic occupancy within the *L3mbtl2* DHS, as revealed by very high read depth digital DNase-seq footprinting analyses. The upper strand (US) and lower strand (LS) DNase cut patterns are shown, revealing a Hoxa9+FoxC1-specific strand cutting imbalance indicative of specific occupancy.

B: Analyses of RNA-seq expression data for *L3mbtl2* together with other constitutively-expressed PRC complex members across the three mouse models. *Tbp* is shown as a housekeeping control.

Further inspection of the *L3mtbl2* locus revealed a Hoxa9+FoxC1-specific DHS present at the second exon which featured two differentially-footprinted sites (Figure 3:25). The first footprint occupied a CTCF site, while the second footprint site spanned one ETS motif and at least one FOX motif. Because *L3mbtl2* encodes a non-canonical PRC repressor protein which deposits repressive chromatin modifications as part of a larger multi-subunit complex[350], we examined the mRNA expression levels of known interacting partners of L3mbtl2. In contrast to robust Hoxa9+FoxC1-dependent upregulation of *L3mbtl2*, the expression of various other subunits of the L3mbtl2 complex were largely invariant between the three different mouse models.

3.12 Activation of the *Meis2* locus in Hoxa9+FoxC1 AMLs

Because upregulation of *Meis2* mRNA was a particularly striking event in Hoxa9+FoxC1 AMLs, we examined the DNase-seq data in more detail at the *Meis2* locus to identify specific DHSs that may account for these changes. Indeed, we found 19 Hoxa9+FoxC1-specific DHSs that were associated with the *Meis2* locus, including a cluster of four promoter-proximal sites and at least 15 distal sites (Figure 3:26).

To gain insight into the regulatory potential of DHSs activated at the *Meis2* locus in a FoxC1-dependent manner, we inspected the underlying DNA sequences for TF binding motifs and evidence of evolutionary conservation. Furthermore, these findings were integrated with genomic occupancy data obtained both from our high-resolution DNase I digital footprinting analyses, publicly available ChIP-seq datasets for FoxC1[351] and CTCF, as well as three active histone modifications H3K4me1/me3 and H3K27ac[151].

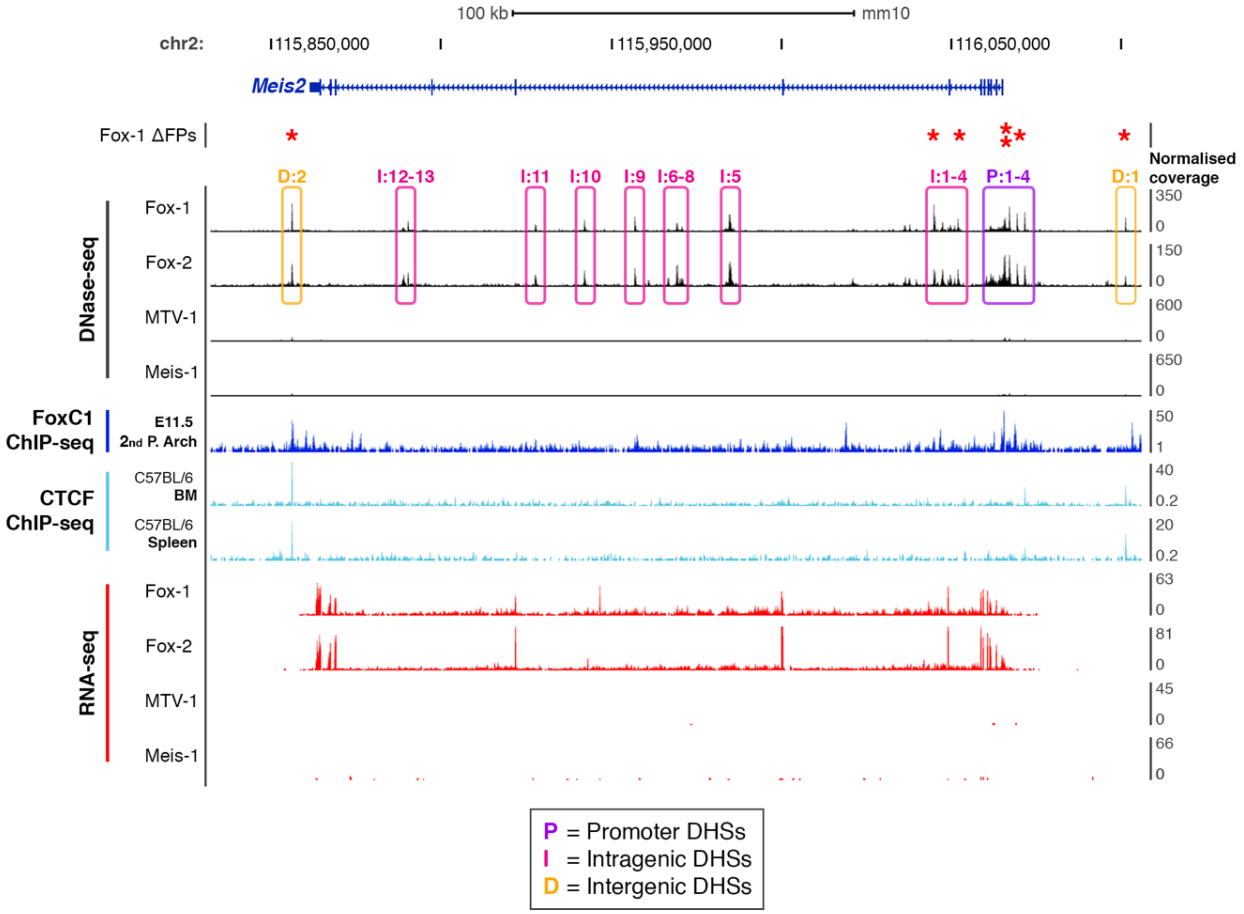


Figure 3:26 **FoxC1 activates the *Meis2* proto-oncogene.**

Top: UCSC genome browser view of DNase-seq data at the *Meis2* locus. Differential DHSs specifically present in Hoxa9+FoxC1 AMLs are indicated in boxes. Footprinted sites within these DHSs that were occupied only in a Hoxa9+FoxC1 AML (Fox-1 ΔFPs) are indicated above the DNase-seq tracks.

Middle: ChIP-seq data from Amin *et al.* (2015) confirms *Meis2* occupancy by FoxC1 in a different cell context. This includes binding at the *Meis2* promoter and a distal regulatory element (red boxes) which are also activated in Hoxa9+FoxC1 AMLs. ENCODE CTCF ChIP-seq data identifies commonly occupied sites in normal bone marrow (BM) and spleen from isogenic mice.

Bottom: UCSC genome browser view of RNA-seq data from the same samples as the DNase-seq at top.

Table 3:2 **Regulatory potential of Hoxa9+FoxC1-specific DHSs present at the *Meis2* locus.**

Occurrence of TF binding motifs and genome occupancy events in Hoxa9+FoxC1-specific DHSs at the *Meis2* locus are indicated. Distal DHSs were qualitatively scored for their gene regulatory potential according to the following criteria:

Digital DNase I Footprint +2 for each footprint

FOX, ETS, RUNX, CTCF motif: +1 for each motif occurrence [within the DHS window].

All other motifs*: +0.5 for each motif occurrence.

CTCF ChIP-seq: +2 for high-confidence binding (> 20 mapped reads) in both BM and Spleen.

FoxC1 ChIP-seq: +1 for weak binding (>12 uniquely mapped reads, presented as +);
+2 for strong binding (> 20 uniquely mapped reads, presented as ++);
(ofs. = ChIP peak summit occurred in proximity to but offset from DHS summit).

Active H3 Modifications: +0.25 for every enrichment (> 5 uniquely mapped reads) in either BM or Spleen.

Genomic Conservation: +1 for high evolutionary conservation (PhyloP placental mammals);
-1 for no conservation.

*only motifs recognised by TFs determined to be expressed in Hoxa9+FoxC1 AMLs (RNA-seq TPM ≥ 1) were considered.

DHS	Distance to 5' UTR	FOX motifs	ETS motifs	RUNX motifs	Other TF motifs	CTCF motifs	CTCF ChIP	FoxC1 ChIP	Active Histone H3 Modifications			DHS Regulatory Potential
									K4me1	K4me3	K27ac	
P:1	- 6.6 kb	-	2	-	SOX (1); Sp1 (1)	-	-	-	++	++	+	
P:2	- 4.5 kb	1	1	-	-	-	-	+ ofs.	++	++	+	
P:3	- 2.1 kb	3	1	-	Sp1 (1)	-	-	+ ofs.	++	++	++	
P:4	- ~650 bp	1	-	-	-	-	-	++	++	++	++	
I:1	+ 12.8 kb	4	-	-	-	-	-	-	++	++	+	**7.25**
I:2	+ 14.8 kb	-	2	-	NFκB (1); Sp1 (1)	-	-	-	++	++	++	**4.5**
I:3	+ 17.5 kb	1	1	-	-	-	-	+ ofs.	++	++	++	**4.5**
I:4	+ 20 kb	1	-	-	STAT (1)	-	-	+	++	++	+	**5.75**
I:5	+ 79 kb	3	1	2	-	-	-	-	++	++	++	**7.5**
I:6	+ 93 kb	-	1	-	-	-	-	-	++	+	++	**1.25**
I:7	+ 93.5 kb	2	-	-	E-box (1)	-	-	-	++	+	++	**3.75**
I:8	+ 94.5 kb	3	-	-	E-box (1); SOX (1)	-	-	-	++	++	++	**5.5**
I:9	+ 108 kb	3	1	-	PBX (1)	-	-	+ ofs.	++	-	-	**6.5**
I:10	+ 123 kb	1	-	1	NFE (1)	-	-	-	++	-	+	**3.25**
I:11	+ 138 kb	2	-	1	-	-	-	+	+	+	++	**5**
I:12	+ 175 kb	1	1	-	-	-	-	-	++	+	-	**2.75**
I:13	+ 177 kb	2	-	-	-	-	-	-	++	++	++	**3.5**
D:1	- 36 kb	2	-	-	E-box (1)	1	++	++ ofs.	++	-	++	**8.5**
D:2	+ 208 kb	-	1	-	E-box (1)	1	++	+ ofs.	++	++	++	**9**

Regulatory Potential	
Low	**< 3**
Mid	**3 - 6**
High	**> 6**

These analyses confirmed enrichment of conserved TF binding motifs – in particular FOX, ETS and/or RUNX sites - at all of the Hoxa9+FoxC1-specific DHSs (Table 3:2). Many of these DHSs were also characterised by an enrichment of three histone marks associated with gene activation, including H3K4me1+3 and H3K27ac. Significantly, we detected the presence of least one differential TF footprint in six of these Hoxa9+FoxC1-specific DHSs, suggesting that these sites may play a direct role in the regulation of *Meis2* in Hoxa9+FoxC1 AMLs.

3.13 Features of the *Meis2* promoter-proximal DHS cluster, P:1-P:4

TF binding motif enrichment, active histone modifications and evidence of TF occupancy were all observed at the promoter-proximal DHS cluster of sites P:1 to P:4 (Figure 3:27). Firstly, P:1 featured a highly conserved FOX binding motif close to the DHS summit, which overlapped with CTCF binding in ChIP-seq data from normal BM and spleen, where *Meis2* is active. Furthermore, P:1 was also enriched for all three activating histone marks in normal BM and spleen. Next, P:2 was distinguished by a specific C/EBP footprint within the DHS and strong enrichment of the active enhancer-associated marks H3K4me1 and H3K27ac. Third, the P:3 DHS was found to be flanked by two conserved FOX binding motifs, one of which is bound by FoxC1 in ChIP-seq data from an embryonic tissue with high *Meis2* expression. Furthermore, the P:3 DHS featured two Hoxa9+FoxC1-specific footprints, one spanning the part of a FOX site and a GC-box like motif, and the other occupying a Sp1 motif. Finally, P:4 exhibited the greatest evolutionary conservation of all the DHSs in this cluster and featured an upstream FOX motif which was also occupied in FoxC1 ChIP-seq data. The entire P:4 DHS was enriched for all three activating histone marks in normal BM and spleen.

Detailed DNA sequence analyses of the P:1-P:4 cluster (Figure 3:28) revealed the presence of other TF binding motifs in addition to FOX sites at these DHSs, including SOX, ETS and Sp1 motifs within these hypersensitive sites.

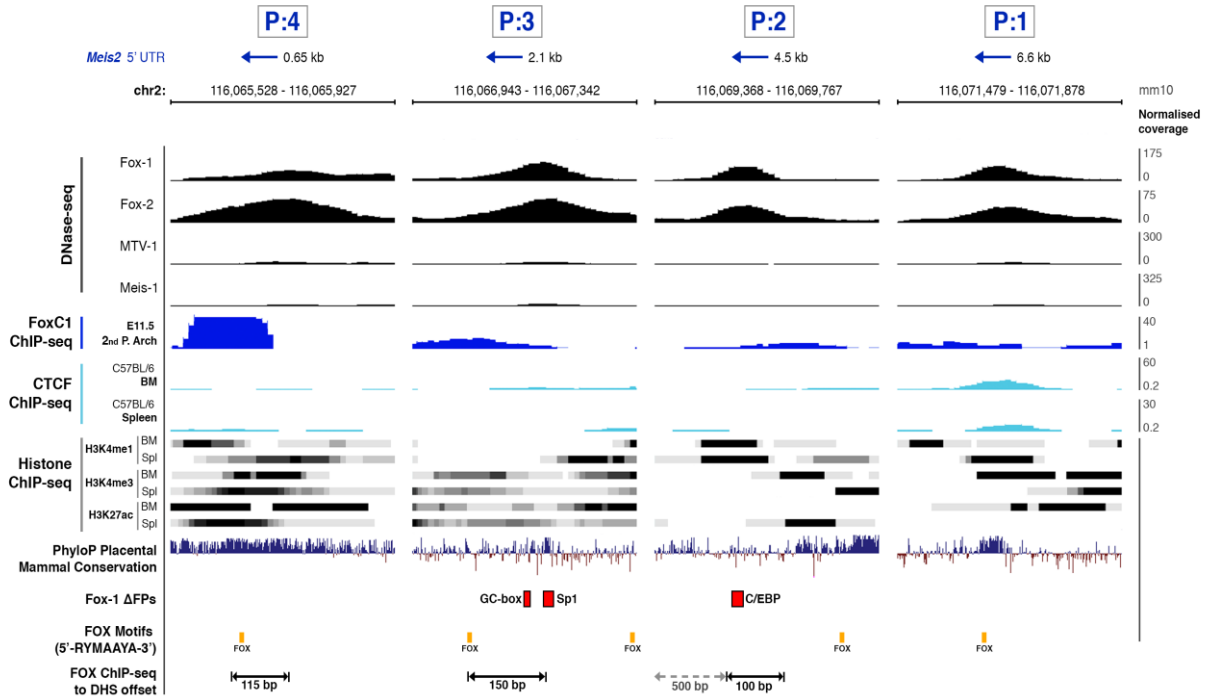


Figure 3:27 **A cluster of *Meis2* promoter-proximal DHSs are specifically activated in Hoxa9+FoxC1 AMLs.**

(Left to right) UCSC genome browser views of four FoxC1-specific promoter-proximal DHSs (**P:1** – **P:4**, in order of increasing genomic distance from the 5' UTR of *Meis2*.

DNase-seq data are shown together with 9 public ChIP-seq datasets (top to bottom = **1-9**), Hoxa9+FoxC1-specific DNase I digital footprints, PhyloP evolutionary conservation data, FOX consensus binding motifs and the measured offset (in base pairs) between Hoxa9+FoxC1-specific DHS and FoxC1 ChIP-seq summits from another study.

(**1**) FoxC1 ChIP-seq data from a study of mouse embryonic 2nd pharangeal arch (2nd P. arch), where *Meis2* is highly active.

(**2-3**) ENCODE CTCF ChIP-seq data from normal C57BL/6 mouse BM and spleen (Spl).

(**4-9**) ENCODE ChIP-seq data also from BM and Spl for three histone modifications associated with gene activation (H3K4me1, H3K4me3 and H3K27ac).

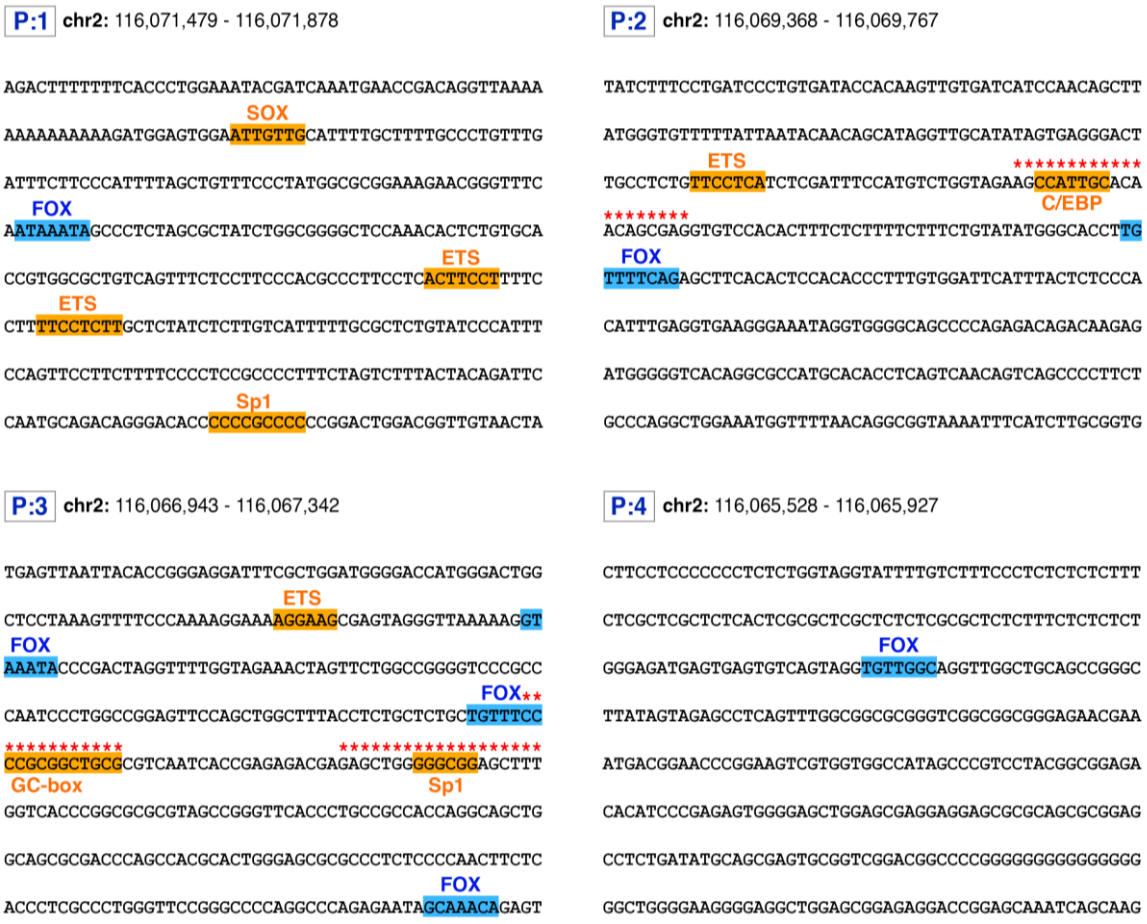


Figure 3:28 ***Meis2* promoter-proximal DHSs are enriched for FOX and other TF binding motifs.**

P:1 – **P:4**: DNA binding motif content for each of the four FoxC1-specific promoter-proximal DHSs described in Table 3:3 and Figure 3:19.

FOX binding motifs are highlighted in blue, and binding sites for other TFs also expressed in Hoxa9+FoxC1 AMLs are highlighted in orange.

Occupied bases in Hoxa9+FoxC1-specific DNase I digital footprints are indicated with red asterixes.

3.14 Features of distal DHSs in Hoxa9+FoxC1 AMLs at the *Meis2* locus

In addition to the promoter-proximal DHS cluster already described, we identified several distal DHSs which were present or substantially upregulated only in the Hoxa9+FoxC1 AMLs that may be related to the expression of *Meis2*. These included two intragenic DHSs, I:1 and I:4, which both featured specific evidence of FOX motif occupancy in Hoxa9+FoxC1 cells (Figure 3:29).

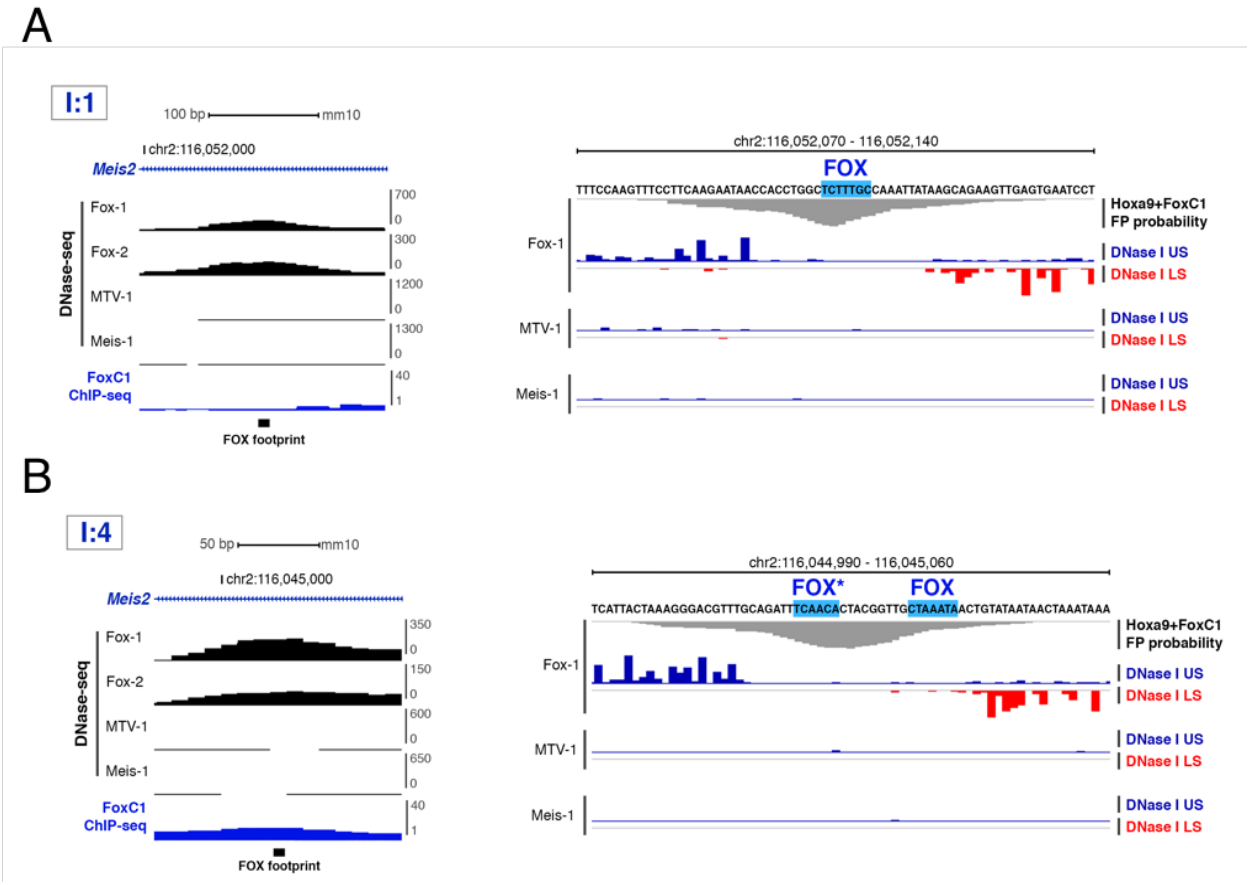


Figure 3:29 **Hoxa9+FoxC1-specific intragenic DHSs at the *Meis2* locus feature high confidence FOX occupancy patterns.**

A: (Left) UCSC genome browser view of the I:4 DHS (+20 kb from the *Meis2* 5' UTR) showing Hoxa9+FoxC1-specific accessibility by DNase-seq, FoxC1 binding by ChIP-seq and FOX motif occupancy using high-resolution DNase-seq digital footprinting analyses.

(Right) detail of occupied FOX sites as revealed by very high read depth digital DNase-seq footprinting analyses. The upper strand (US) and lower strand (LS) DNase cut patterns are shown, revealing a Hoxa9+FoxC1-specific strand cutting imbalance indicative of specific occupancy. The site indicated FOX* is a non-canonical FOX motif which can nevertheless be bound by FOX proteins.

B: (Left) UCSC genome browser view of the I:1 DHS (+12.8 kb from the *Meis2* 5' UTR) showing Hoxa9+FoxC1 specific accessibility and TF occupancy through analyses of high-resolution DNase-seq data.

(Right) detail of occupied FOX site revealed by high read depth digital DNase-seq footprinting analyses as in (A).

Either side of the *Meis2* gene, we identified two Hoxa9+FoxC1-specific DHSs, D:1 and D:2, which both featured evidence of CTCF binding in normal BM and spleen[151,335] as

revealed by analyses of ENCODE ChIP-seq data. Additionally, analyses of the DNA sequences in site D:1 revealed the presence of the sequence 5'-GTGCAGTACC-3' immediately upstream of the main CTCF consensus, making a conserved CTCF-composite site that is predicted to have insulator activity (Figure 3:30)[352]. Critically, our DNase I digital footprinting analyses revealed that the composite CTCF insulator site at D:1 was occupied in a Hoxa9+FoxC1-specific manner. Furthermore, both D:1 and D:2, together with several other distal DHSs, were found to be occupied by FoxC1 in ChIP-seq data from another study as observed at the promoter-proximal cluster of DHSs (Table 3:2).

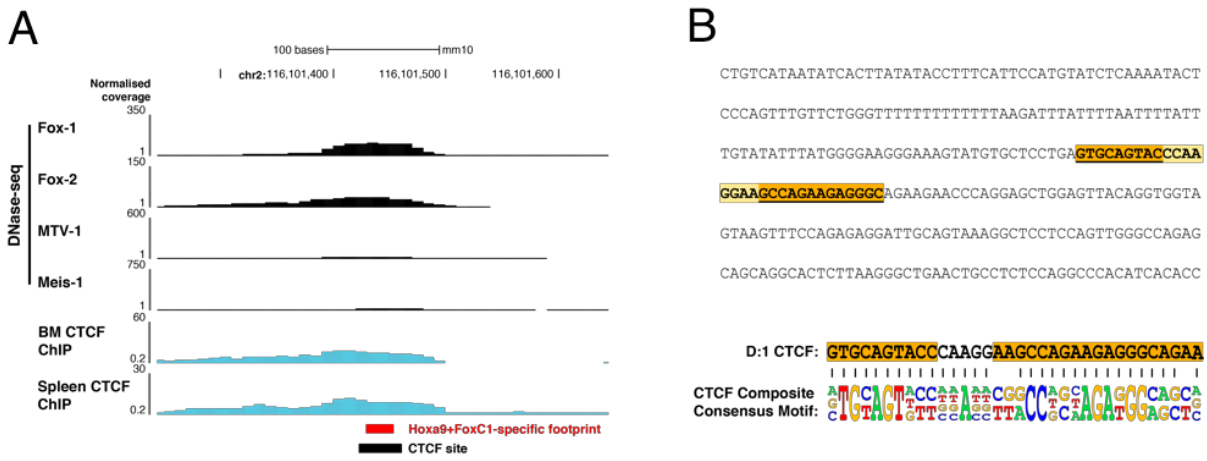


Figure 3:30 **The *Meis2* D:1 DHS contains a conserved CTCF composite site with predicted insulator activity.**

A: UCSC genome browser view of the *Meis2* D:1 DHS featuring evidence of CTCF occupancy in normal mouse BM and spleen as shown by ENCODE ChIP-seq data. A Hoxa9+FoxC1 AML-specific footprint overlapping a CTCF site is indicated below.

B: (Top) genomic sequence detail of a putative CTCF composite site differentially occupied in a Hoxa9+FoxC1 AML, as evidenced by DNase I digital footprinting. (Bottom) conservation of the D:1 site (core sequences highlighted) with the consensus CTCF composite motif, with conserved bases connected by vertical dashes.

Interestingly, we observed FoxC1 binding at the *Meis2* locus often occurred close to Hoxa9+FoxC1-activated DHSs, but outside of the DNase I-hypersensitive region, consistent with nucleosomal binding (Figure 3:31). Offset binding of FoxC1 with respect

to nearby DHSs was an unexpected event, given that ChIP-seq signal for most TFs commonly overlap with hypersensitive regions, but this was a recurring event we also noticed in directly comparable DNase-seq and ChIP-seq datasets generated in FUJIOKA cells, which are discussed in Chapter 4. Finally, as our digital DNase I footprinting algorithm only detects TF occupancy patterns as they exist in accessible chromatin, putative offset binding of FoxC1 may also explain why we unexpectedly did not observe a significant enrichment of occupied FOX motifs in our global footprinting analyses (Figure 3:14).

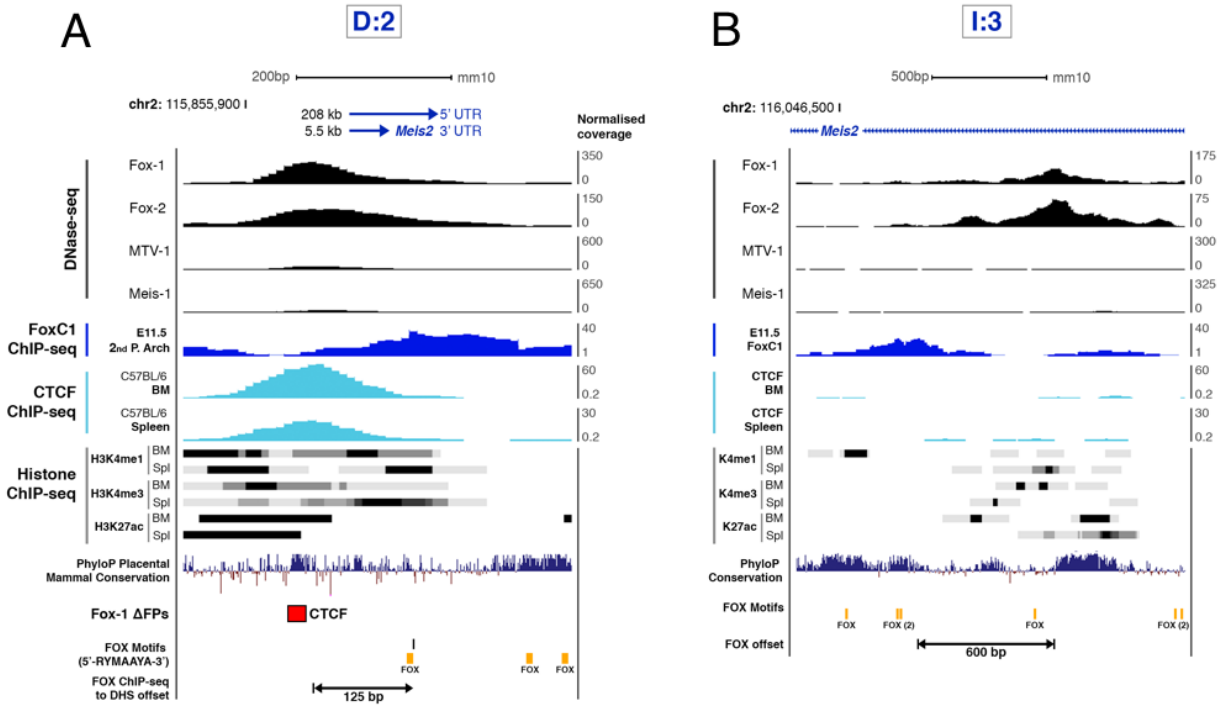


Figure 3:31 **Offset binding of FoxC1 relative to Hoxa9+FoxC1 AML specific DHSs, and other features of elements activated at the *Meis2* locus.**

UCSC genome browser views of two FoxC1-specific DHSs together with public ChIP-seq datasets, evolutionary conservation, FOX consensus binding motifs and offset FoxC1 binding in ChIP-seq from another study.

A: The +208 kb intergenic DHS (**D:2**) features a Hoxa9+FoxC1 specific CTCF footprint, together with CTCF peaks in normal BM and spleen ChIP-seq data. This region is also distinguished by enrichment of activating histone marks H3K4me1+3 and H3K27ac. Three highly-conserved FOX binding motifs are present at this site and are occupied by FoxC1 in an unrelated tissue. FoxC1 ChIP-seq signal is enriched outside the DHS, with the FoxC1 ChIP-seq summit in this window offset from the D:2 DHS summit by ~125 bp.

B: The +17.5 kb intragenic DHS (**I:3**) features modest enrichment of all three histone marks, but particularly the H3K27ac modification associated with active enhancer regions. This window features 6 FOX consensus binding sites, 5 of which are highly conserved. FoxC1 ChIP-seq data reveals this site to be occupied in an unrelated tissue, with maximal binding positioned ~600 bp downstream from the I:3 DHS summit.

3.15 Genes deregulated by Hoxa9 and FoxC1 in mouse AML are commonly deregulated in FLT3-ITD human AMLs

Given that *FOXC1* was previously reported as a target gene of FLT3-ITD in human AML patients, where it was associated with specific changes in gene expression, chromatin accessibility and TF occupancy patterns[271], we sought to investigate commonalities between the genes deregulated in FoxC1-dependent mouse AMLs and human FLT3-ITD AML patients.

A comparison of differentially expressed mRNAs in the Hoxa9+FoxC1 AMLs (log_2 FC ≥ 1.0, p <0.05) revealed substantial overlap with the FLT3-ITD AML target genes identified in Cauchy *et al.* (2015) (Figure 3:32). Firstly, 11 of the 134 genes specifically upregulated in FLT3-ITD AML patients were also significantly upregulated in Hoxa9+FoxC1 mice, including *Prickle1* (Table 3:3), previously connected with leukaemogenesis[353]. Further inspection of the chromatin landscape surrounding commonly upregulated genes revealed the presence of Hoxa9+FoxC1-specific DHSs and TF footprints, for example at *Spry2* and *Prickle1* (Figure 3:33).

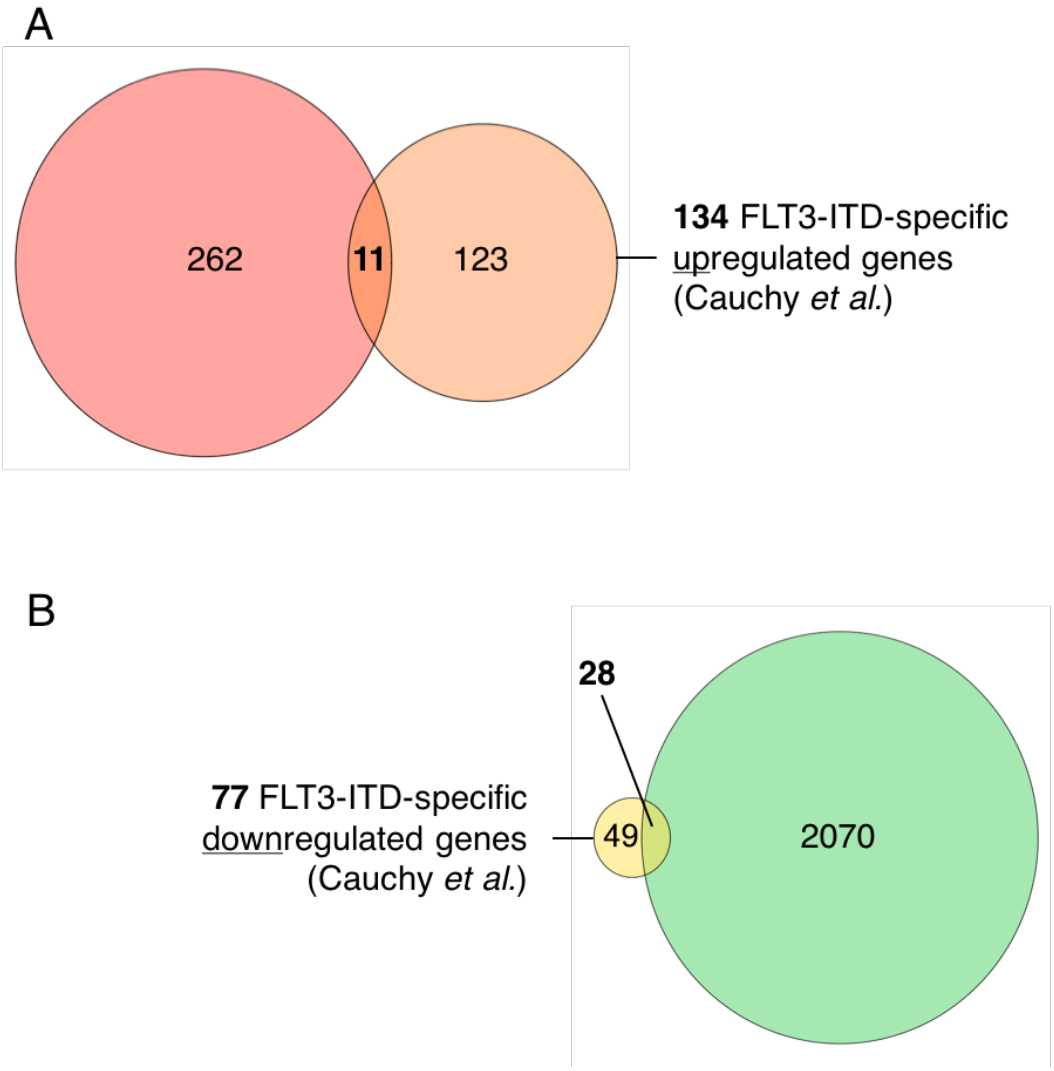


Figure 3:32 **FoxC1 activates a modest number of target genes in mouse AMLs that are common with human FLT3-ITD+ AML patient samples.**

Venn diagrams showing the overlap between Hoxa9+FoxC1 specific targets in mouse AMLs (compared to Hoxa9 alone) and FLT3-ITD-specific upregulated and downregulated genes (**A** and **B** respectively) in primary human AML samples from microarray analyses in Cauchy *et al.* (2015).

Expression thresholds for significantly upregulated and downregulated genes in the mouse AML RNA-seq data were set to comparable levels to the Cauchy *et al.* (2015) dataset (log2 FC >1, *adj. p* <0.05).

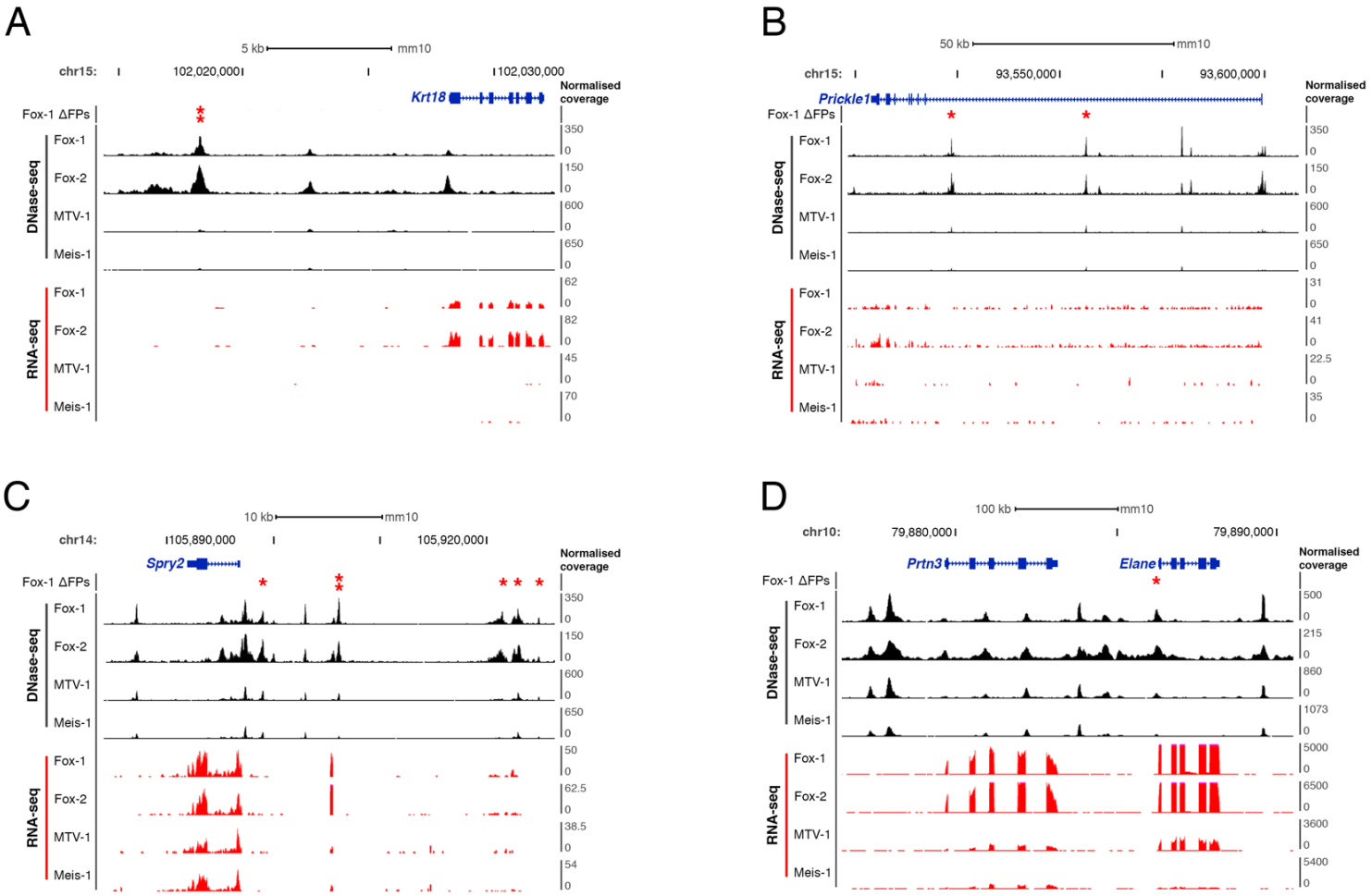


Figure 3:33 **Evidence of Hoxa9+FoxC1-specific regulation of commonly upregulated genes in human FLT3-ITD AMLs**

A-D: UCSC genome browser view of DNase-seq and RNA-seq data for four loci (*Krt18*, *Prickle1*, *Spry2* and the *Prtn3/Elane* cluster) which are commonly upregulated both in Hoxa9+FoxC1 AMLs and the Cauchy *et al.* (2015) FLT3-ITD-specific gene list. DHSs which feature specific footprints in Hoxa9+FoxC1 cells are indicated with red asterixes.

Table 3:3 **Hoxa9+FoxC1 AMLs share commonly activated and repressed genes with primary FLT3-ITD+ AML patient samples.**

Analysis of RNA-seq mRNA expression fold changes of Hoxa9+Foxc1 relative to Hoxa9+MMTV for significantly deregulated genes that are common to those of human FLT3-ITD AML patient samples from Cauchy *et al.* (2015). Genes coloured in grey are significantly deregulated, but fall below the nominal threshold for substantially changing genes (log2 FC > 1.0).

All commonly upregulated genes (mFoxC1 AML and FLT3-ITD genes)

Mouse gene	Human orthologue	Foxc1/MMTV log2 FC	*p* value
Foxc1	*FOXC1*	3.97	3.49E-05
Ms4a3	*MS4A3*	4.08	1.14E-02
Krt18	*KRT18*	3.28	1.03E-03
Elane	*ELANE*	3.23	1.01E-04
Prtn3	*PRTN3*	2.60	6.83E-05
Spns3	*SPNS3*	2.03	2.36E-04
Rph3al	*RPH3AL*	1.47	3.29E-03
Prickle1	*PRICKLE1*	1.32	1.30E-02
Hoxb4	*HOXB4*	1.31	8.31E-02
Ctsg	*CTSG*	1.27	2.33E-02
Fam83a	*FAM83A*	1.19	5.57E-03
Snap47	*SNAP47*	0.94	1.53E-02
Spry2	*SPRY2*	0.67	4.46E-02

All commonly downregulated genes (Hoxa9+FoxC1 and FLT3-ITD genes)

Mouse gene	Human orthologue	Foxc1/MMTV log2 FC	*p* value
Il10ra	*IL10RA*	-4.78	2.54E-06
Trps1	*TRPS1*	-4.30	3.36E-06
Ifitm3	*IFTM3*	-4.24	6.42E-04
Fcgrt	*FCGRT*	-4.18	6.17E-05
Mx1	*MX1*	-3.94	7.00E-06
Cd48	*CD48*	-3.77	1.58E-06
Dpysl2	*DPYSL2*	-3.33	7.70E-06
Aim1	*AIM1*	-3.29	9.97E-05
Gng2	*GNG2*	-3.22	3.33E-05
Plcb1	*PLCB1*	-3.19	7.44E-05
Mdfic	*MDFIC*	-2.97	1.49E-05
Kcna3	*KCNA3*	-2.69	1.80E-05
Rcan3	*RCAN3*	-2.19	5.09E-05
Zeb2	*ZEB2*	-2.02	6.30E-04
Pde4b	*PDE4B*	-2.00	1.22E-03
H2-DMb1	*HLA-DMB*	-1.87	2.84E-04
H2-Aa	*HLA-DQA1*	-1.77	1.29E-02
Dock10	*DOCK10*	-1.66	4.61E-04
Foxo1	*FOXO1*	-1.47	2.22E-03
Hdgfrp3	*HDGFR3*	-1.42	2.61E-02
Xpa	*XPA*	-1.31	2.32E-03
H2-DMb2	*HLA-DMB*	-1.29	1.36E-02
Prkd3	*PRKD3*	-1.22	2.81E-03
Sell	*SELL*	-1.17	5.53E-03
Xaf1	*XAF1*	-1.14	4.44E-03
Rgs10	*RGS10*	-1.11	6.97E-03
Rab43	*RAB43*	-1.05	1.85E-02
Akt3	*AKT3*	-1.04	5.87E-03
Setbp1	*SETBP1*	-0.80	4.85E-02

Next, 28 of the 77 genes specifically downregulated in FLT3-ITD human AMLs were also downregulated in the Hoxa9+FoxC1 AML samples, including the pro-apoptotic factor *Xaf1*[354] and murine orthologues of the human major histocompatibility (MHC) genes *HLA-DMB* and *HLA-DQA1*.

Finally, as downregulation of MHC genes was a significant feature of human FLT3-ITD AMLs, and MHC genes diverge significantly between species[355], we investigated the full complement of MHC genes in FoxC1-dependent mouse AMLs that were significantly deregulated. Three of the MHC genes downregulated in the human AMLs were also down-regulated in the mouse Hoxa9+Meis1 AML, plus we observed substantial downregulation of a further 12 MHC genes in Hoxa9+FoxC1 AMLs, 11 of which had conserved human orthologues (Table 3:4).

Table 3:4 **Widespread downregulation of major histocompatibility complex genes is a feature of the Hoxa9+FoxC1 leukaemia gene expression signature.**

RNA-seq data were analysed as in Table 3:2 and 3:3, with any human orthologues commonly downregulated in FLT3-ITD AML patient samples from Cauchy *et al.* (2015) excluded from this analysis to avoid duplication.

Additional MHC genes downregulated in Hoxa9+FoxC1 AMLs alone

Mouse gene	Human orthologue	Foxc1/MMTV log2 FC	*p* value
H2-Q1	*HLA-A*	-2.76	7.82E-02
H2-M3	*HLA-G*	-2.22	1.87E-03
H2-DMa	*HLA-DMA*	-2.02	1.18E-03
H2-Ob	*HLA-DOB*	-1.86	1.89E-02
H2-T23	*HLA-E*	-1.75	8.12E-04
H2-Eb1	*HLA-DRB5*	-1.69	9.72E-03
H2-Q4	*HLA-A*	-1.62	1.81E-02
H2-Ab1	*HLA-DQB1*	-1.53	1.22E-02
H2-T22	n/a	-1.40	1.33E-02
H2-D1	*HLA-A*	-1.31	1.37E-02
H2-Q7	*HLA-A*	-1.23	4.62E-02
H2-Q2	*HLA-A*	-1.04	3.74E-01

Chapter 3: Discussion

3.16 DNase-seq analyses support existing gene expression data to indicate a functional relationship between *FOXC1* and *HOXA9* in primary human AML samples and cell lines

Previous studies had ascertained a positive correlation between mRNA expression of *FOXC1* and *HOXA9* in human AMLs[300], although it remained unclear whether there was a direct relationship between the way these two genes were regulated in AML. Through DNase-seq profiling of primary human AML samples and cell lines with varying levels of *HOXA9* and *FOXC1* mRNA expression (Figure 3:2 and Figure 3:3), we observed that the entire *HOXA* chromatin domain featured increased accessibility in samples with higher *FOXC1* expression, strengthening the possibility of a functional relationship between the regulation of these two TF genes.

3.17 Hoxa9+FoxC1 donor cells carry an abnormal myeloid phenotype consistent with the development of AML

Our FACS immunophenotyping analyses indicated that all the mouse AML samples studied contained a sufficiently high proportion of CD45.1$^+$ donor leukaemic cells to permit detailed downstream molecular analyses (Figure 3:5). These data confirmed that the Hoxa9+FoxC1-dependent AMLs infiltrated recipient spleens most extensively, closely followed by Hoxa9 with Meis1, a widely-documented oncogenic partner of Hoxa9[330–332]. Comparatively lower levels of engraftment for cells transduced with Hoxa9+MMTV is also consistent with findings elsewhere that the forced expression of Hoxa9 alone is not sufficient to induce frank leukaemia in mice in the absence of additional genetic factors, instead causing a myeloproliferative disorder[356].

We also observed that the CD45.1$^+$ donor fractions of all samples studied were enriched for the CD11b and Gr1 myeloid surface markers, but that this was most striking in

Hoxa9+FoxC1 AMLs. Considering that a healthy mouse spleen is comprised of ~80% lymphoid cells[357], these findings alone are indicative of a significant disruption to normal haematopoiesis. In another study of an MLL-AF9-driven mouse model of AML, 90% of LSCs detected in recipient spleens carried the $Gr1^{+}Cd11b^{+}$ phenotype, suggesting that the Hoxa9+FoxC1 samples were also enriched for LSCs and AML blast cells[358]. Finally, $CD45.1^{+}$ donor cells also featured abnormally high levels of the primitive B cell marker CD43, which is normally lost from all but a minority population of peripheral B cells that reside in the spleen[359]. Critically, our FACS observations are consistent with previously published analyses of BM samples from the same mice[300].

Due to the substantial enrichment of abnormal $CD45.1^{+}$ myeloid-like cells within these samples, we did not consider that the contaminating host cell populations - which were also much more heterogeneous in nature - were likely to impede subsequent genome-wide analyses. These assumptions were corroborated by our RNA-seq analyses (Figure 3:19), which demonstrated substantial Hoxa9+FoxC1-dependent downregulation of mature myeloid marker genes including *Cd14* and three chemokine (C-C) motif ligand genes *Ccl3*, *Ccl4* and *Ccl6*[360–362]. Furthermore, the striking concordance between DNase-seq and RNA-seq profiles of Hoxa9+MMTV and Hoxa9+Meis1 AMLs (Figure 3:7 and Figure 3:17), despite varying levels of $CD45.2^{+}$ cell composition between the two groups, are a strong indication that the $CD45.2^{+}$ cells make a marginal contribution to our analyses.

The Hoxa9+FoxC1-specific depletion of DHSs containing IRF binding motifs, in parallel with the downregulation of five *Irf* transcription factor genes, are strong indications that downregulation of normal myeloid-associated gene expression patterns is a defining feature of Hoxa9+FoxC1-dependent leukaemogenesis (Figure 3:11and Figure 3:19)[336,340,341,363]. Trends observed by manual inspection of differentially-expressed gene

lists in the FoxC1-dependent AMLs were supported by KEGG pathway analyses (Table 3:1), which revealed significant downregulation of numerous genes that are active in normal immune cells, including Toll-like receptor, TNF and NF-κB signalling[364,365].

3.18 FoxC1 promotes widespread reprogramming of accessible chromatin and gene expression in AML via the deregulation of proto-oncogene TFs

In this study, we sought to expand upon previous phenotypic analyses of FoxC1 in AML by identifying the target genes and *cis*-regulatory elements differentially regulated in FoxC1-dependent AMLs. We were successful in demonstrating a specific and robust chromatin signature activated in Hoxa9+FoxC1-dependent AMLs (Figure 3:7, Figure 3:8 and Figure 3:9), which strongly correlated with changes in gene expression (Figure 3:16 and Figure 3:17).

We dissected the accessible chromatin signature of Hoxa9+FoxC1 AMLs further by investigating the TF motif composition of the most specific DHSs in these cells, and identified a FoxC1-dependent enrichment of C/EBP, MEIS and MYB binding motifs. These analyses were corroborated by motif co-localisation analyses, revealing that these DHSs were defined by an enrichment of HOX, FOX, C/EBP and MYB motifs (Figure 3:12). Collectively, these global analyses implicated these factors as principal regulators of the Hoxa9+FoxC1-dependent transcription network. This was in contrast to an underlying motif binding signature of PU1:IRF, HOX, RUNX and AP-1 sites defining Hoxa9+MMTV-specific DHSs, which was lost from the Hoxa9+FoxC1 AML cells.

The Hoxa9+FoxC1-specific enrichment of *de novo* C/EBP motifs (Figure 3:10 and Figure 3:11), occupied C/EBP sites (Figure 3:13), and striking upregulation of *Cebpe* (Figure 3:21), strongly suggests that deregulation of C/EBP proteins is a significant event in FoxC1-dependent transformation. Although upregulation of C/EBP family proteins,

including *Cebpe* in mice, is typically associated with myeloid progenitor expansion and eventual differentiation, the timing and level of C/EBP expression is critical. An essential role for C/EBPε in granulopoesis has been demonstrated in mouse knockout studies[366], although there is currently no evidence in mice that overexpression of wild-type C/EBPε promotes leukaemogenesis. However, *CEBPE* is the target of chromosomal rearrangements of *IGH* in human lymphoid malignancies, where the wild-type *CEBPE* allele is translocated close to an enhancer site normally responsible for *IGH* activation, leading to upregulation of *CEBPE* with leukaemogeneic consequences[367]. Furthermore, there is evidence that overexpression of other C/EBP family members can have oncogenic consequences, including C/EBPδ in mouse AML and C/EBPβ in human solid tumours[368–370]. For example, in mouse AMLs, C/EBPδ deregulation is accompanied by the suppression of C/EBPα expression, promoting a myeloid differentiation block. In our own analyses, we observed only a modest downregulation of *Cebpa* expression in Hoxa9+FoxC1 AMLs as compared to Hoxa9+MMTV ($\log_2$ FC = 1.8, p = 0.04), but this did not survive correction for multiple testing (adj. p = 0.11). We speculate that overexpression of C/EBPε might have functional consequences relevant to the development of AML, although the downstream targets are not yet clear (see 3.20). We also consider the possibility that C/ebpε upregulation in the FoxC1-dependent AML mouse model is simply a negative feedback event to resist downregulation of the myeloid expression programme in these samples. However, given the striking overrepresentation of CEBP motifs present in FoxC1-specific DHSs, including those associated with both up- and downregulated genes, we consider this unlikely.

Our observation that Hoxa9+FoxC1 AMLs featured upregulation of *Myb* and an enrichment of DHSs containing MYB binding sites (Figure 3:10 and Figure 3:23) may be a significant aspect of the gene regulatory network activated by Hoxa9 and FoxC1.

Firstly, in addition to essential roles in regulating the self-renewal of normal HSPCs[371,371], *MYB* is a defined oncogene in both solid malignancies and leukaemias, where it is frequently amplified or overexpressed[372,373]. In mice, *Myb* is a validated target gene of Hoxa9 and Meis1 in AML, where *Myb* upregulation directly contributes to leukaemic transformation[374]. In Hoxa9+FoxC1 AMLs, the substantial FoxC1-dependent upregulation of *Meis2* (Figure 3:26), which can exhibit functional redundancy with Meis1, may be important in this regard[331,332].

Critically, the considerable overlap between the accessible chromatin and gene expression profiles of Hoxa9+Meis1 AMLs and samples transduced with Hoxa9+MMTV is consistent with the previously defined role of Meis1 in stabilising Hoxa9 at target loci, rather than retargeting Hoxa9 binding to elsewhere[375]. Nevertheless, the Hoxa9+FoxC1-specific chromatin and gene expression signature we defined remained remarkably similar whether compared to Hoxa9+Meis1 AMLs or Hoxa9+MMTV cells, strongly indicating that FoxC1 plays a significant role in the leukaemic transformation of these cells at both the epigenetic and transcriptional level.

Finally, the above indications of Hoxa9+FoxC1-dependent deregulation of C/EBP and Myb proteins were supported further by motif co-localisation analyses (Figure 3:12). These revealed that the most specific DHSs in Hoxa9+FoxC1 AMLs were not only defined by an expected co-localisation of HOX and FOX motifs, but that these motifs were also co-associated with C/EBP and MYB sites. Collectively, these data provide a strong indication that the gene regulation network established in Hoxa9+FoxC1 AMLs is defined by aberrant co-option of C/EBP and Myb activity.

3.19 Activation of *Meis2* is a significant event in Hoxa9+FoxC1 AMLs and is associated with widespread changes in chromatin accessibility across the entire *Meis2* domain

To our surprise, we observed a specific enrichment of MEIS motifs within DHSs specifically activated in Hoxa9+FoxC1 AMLs (Figure 3:11) that were not present even in the Hoxa9+Meis1 AMLs. This was likely related to the striking FoxC1-dependent upregulation of *Meis2* (Figure 3:20), as other *Meis* genes were not expressed in Hoxa9+FoxC1 AMLs and *Meis2* was not expressed in the samples lacking FoxC1 expression (Figure 3:21).

Through detailed analyses of DHSs associated with the *Meis2* locus that were activated in Hoxa9+FoxC1 AMLs alone (Figure 3:26), we found compelling evidence of at least 19 promoter-proximal and distal *cis*-regulatory elements present at the *Meis2* locus in Hoxa9+FoxC1 AMLs but not the other samples. These included two distal DHSs (D:1 and D:2) that bounded the 5' and 3' ends of *Meis2* respectively, and were occupied by CTCF in normal mouse BM and spleen samples (Figure 3:26 and Table 3:2). The sequence and chromatin features of these two sites are likely of particular significance in *Meis2* activation. Firstly, because site D:1 contained a highly conserved composite CTCF site demonstrated in other studies to have insulator activity[352] (Figure 3:30), this DHS possibly acts as the 5' boundary of a ~250 kb chromatin domain activated during *Meis2* upregulation. This prospect is strengthened further by the presence of a Hoxa9+FoxC1 AML-specific CTCF footprint in our high-resolution DNase-seq digital footprinting analyses.

Secondly, site D:2 features a canonical CTCF site occupied in a Hoxa9+FoxC1-specific manner by DNase I digital footprinting, as well as in normal mouse BM and spleen CTCF ChIP-seq, which also express *Meis2*[335]. Robust enrichment of the enhancer mark

H3K27ac[376] and localised binding of FoxC1 in ChIP-seq data, strongly suggest that this DHS could function as a long-range enhancer which may interact with the *Meis2* promoter via looping during gene activation. Indeed, evidence for co-operativity between CTCF and another FOX protein, FoxA1, was previously described in ER-dependent breast cancer cells at cell-specific enhancer elements[377]. Distinct roles for CTCF as a mediator of longer-range enhancer-promoter interactions are emerging[378], further increasing the possibility that site D:2 might play a role in the direct activation of *Meis2* expression.

Analyses of a public FoxC1 ChIP-seq dataset from embyronic mesoderm, where *Meis2* is highly active[351], revealed examples of offset FoxC1 binding in ChIP-seq data with respect to the D:2 DHS and other specific sites at the *Meis2* locus. We speculate that this unusual mode of binding to nucleosomal sites may be evidence of FoxC1 pioneer activity. In addition to potential evidence of nucleosomal FoxC1 binding at the *Meis2* locus, we identified at least two examples of FOX motif occupancy at Hoxa9+FoxC1-specific DHS summits (Figure 3:29), consistent with a more traditional model of TF binding at pre-existing open chromatin sites[2]. Although our mechanistic analyses of FoxC1 binding to chromatin are incomplete, studies of other FOX proteins show similar bimodal binding properties, supporting the possibility that FoxC1 may eventually be classified as a pioneer factor[140,379].

Many of the *Meis2* DHSs described in this study existed in regions of high DNA sequence homology, suggesting that these *cis*-regulatory elements may be conserved and indeed involved in *MEIS2* regulation in humans. This is of relevance to the biology of AML: in a study of human AMLs with *RUNX1-ETO* mutations, MEIS2 was demonstrated to have leukaemogenic properties by associating with the DNA-binding domain of the RUNX1-ETO oncoprotein, impairing the ability of RUNX1-ETO to repress target

oncogenes[380]. Future analyses could compare the activated sites at *Meis2* in Hoxa9+FoxC1 AMLs with primary human AML samples to identify critical regulators of *MEIS2* activation in this context.

Collectively, these data strongly indicate that *Meis2* may be a direct target gene of FoxC1 in Hoxa9+FoxC1 AMLs, a phenomenon that would be directly relevant to human disease. Additional experiments are required to validate mechanistic aspects of FoxC1-dependent activation of *Meis2*, which are explored in 3:23.

3.20 Widespread repression of differentiation-specific gene expression may be related to specific upregulation of the Polycomb regulator *L3mbtl2* in Hoxa9+FoxC1 AMLs

Given that previous biochemical analyses strongly suggested that FoxC1 is a transcriptional activator and not a repressor[381,382], we searched for alternative mechanisms whereby FoxC1 could indirectly promote the widespread downregulation of mRNA expression we observed in Hoxa9+FoxC1 AMLs (Figure 3:16 - Figure 3:19).

Through integrative analyses of RNA-seq and DNase-seq DHS and high-resolution digital footprinting data, we identified *L3mtbl2* as a substantially upregulated target of Hoxa9 and FoxC1 (Figure 3:20 and Figure 3:24). *In vitro* studies have demonstrated that L3MBTL2 alone is sufficient to compact nucleosomal arrays, consistent with a role in transcriptional repression[344]. Furthermore, an essential role for L3mbtl2 in development was demonstrated using knockout studies: *L3mbtl2*-null mouse embryos are inviable, while *L3mtbl2*$^{-/-}$ ES cells are defective in both proliferation and differentiation[350].

In addition to intrinsic nucleosome compacting ability, L3mbtl2 functions as part of a larger PRC1-related complex (PRC1-like 4, or PRC1L4). In mouse ES cells, L3mbtl2 targets PRC1L4 to differentiation-specific genes for repression, promoting proliferation and maintaining pluripotency[350]. Indications of a potential mechanism for this effect were

revealed in another study, which demonstrated that L3MBTL2 has intrinsic ubiquitin E3 ligase activity[344]. Indeed, L3MBTL2 catalyses deposition of H2A lysine 119 monoubiquitination (H2AK119ub) to target genes for repression via other members of this PRC1L4 complex. Significantly, we confirmed that several other genes encoding subunits of the PRC1L4 complex were expressed in Hoxa9+FoxC1 AMLs, including *Ring1*, *Pcgf6, Cbx3* and *E2f6* (Figure 3:25).

Collectively, our analyses indicate that L3mbtl2, perhaps as part of the PRC1L4 complex, is co-opted in Hoxa9+FoxC1 AMLs, where its activity may contribute to the observed myeloid differentiation block by causing widespread repression of differentiation-specific genes. This potential mechanism is consistent with observations in other leukaemias including a mouse model of MLL-AF9 AML, where PRC-dependent activity was found to suppress the expression of similar differentiation-specific genes[383].

Although our observations raise the possibility that misregulation of Polycomb complexes might contribute to the repression of differentiation-specific genes in AMLs with high Hoxa9+FoxC1 expression, alternative mechanisms could equally account for these consequences. It is possible, for example, that FoxC1 could simply sequester or cause retargeting of other TFs away from differentiation-specific genes via one of its transactivation domains[381]. However, given that FoxC1 has only been characterised as a transcriptional activator to date, this possibility is perhaps less likely[299,381,384]. Future experiments aim to address these outstanding questions, and are discussed in section 3:23.

3.21 Commonalities and differences between the gene expression profiles of Hoxa9+FoxC1 mouse AMLs and primary human FLT3-ITD AMLs

Because *FOXC1* was previously reported to be upregulated in FLT3-ITD AMLs[271], we sought to assess how the gene expression profiles of Hoxa9+FoxC1 AMLs were related to earlier observations made via analyses of primary FLT3-ITD AML samples (Figure 3:32 and Table 3:3). Indeed, we found evidence of 39 genes that were commonly deregulated in both the Hoxa9+FoxC1 mouse AMLs and the Cauchy *et al.* (2015) FLT3-ITD-specific gene list. 12 of these genes were upregulated in Hoxa9+FoxC1 AMLs, including *Elane* and the established AML-associated antigen *Prtn3*, which both exist as a cluster of upregulated genes in human FLT3-ITD AMLs[271] as well as in this mouse model (Figure 3:33). Furthermore, *Prickle1* and *Spry2* were two additional commonly upregulated genes with evidence of FoxC1-dependent deregulation. *PRICKLE1* has previously been associated with poorer clinical prognosis in lymphoid malignancies[353], whereas a role for *SPRY2* in leukaemia emerged in a recent study published earlier this year. However, we cannot yet conclude whether these effects are direct or indirect consequences of FoxC1 overexpression, given a current absence of functional validation. Ongoing work (3:23) aims to address these outstanding questions.

In addition to the upregulated genes discussed, our analyses revealed 28 commonly downregulated genes between Hoxa9+FoxC1 mouse AMLs and the human FLT3-ITD AML-specific gene list. These included *Xaf1*, a pro-apoptotic protein which is often downregulated in malignancies including AML[354,385], together with three murine orthologues of human HLA genes: *H2-DMb1*, *H2-Aa* and *H2-DMb2*. The latter trio of downregulated genes is particularly interesting, given the primary FLT3-ITD AML samples analysed in Cauchy *et al.* (2015) featured a broad reduction in the expression of MHC genes. Furthermore, we found substantial Hoxa9+FoxC1-dependent

downregulation of an additional 12 MHC genes (Table 3:4) that were not commonly downregulated in the human FLT3-ITD AML gene list. These findings indicate that Hoxa9 and FoxC1 may potentially be involved in the repression of genes associated with immune recognition. This may be clinically significant, since immune evasion is a characterised hallmark of AML and is an area of active interest for the development of novel therapeutic agents[386].

3.22 Conclusions

Through genome-wide interrogation of chromatin accessibility, TF footprinting and gene expression, we have successfully begun to identify a gene regulatory network activated by ectopic expression of FoxC1 in mouse AMLs (Figure 3:34). We have described a robust pattern of gene expression activated by Hoxa9+FoxC1, including upregulation of *Cebpe* and the proto-oncogenes *Myb* and *Meis2*, and note that the chromatin signature associated with these global changes in gene expression are defined by DHSs enriched for sites that could be bound by these factors. Furthermore, FoxC1 collaborates with Hoxa9 in this model to repress the myeloid gene expression programme, including eight IRF and KLF-family TF genes, together with the downregulation of DHSs containing motifs recognised by these factors. We speculate that the myeloid differentiation block in Hoxa9+FoxC1 AML cells may in part be driven by L3mbtl2-mediated repression of differentiation-specific genes, which we observed to be strikingly upregulated in these cells.

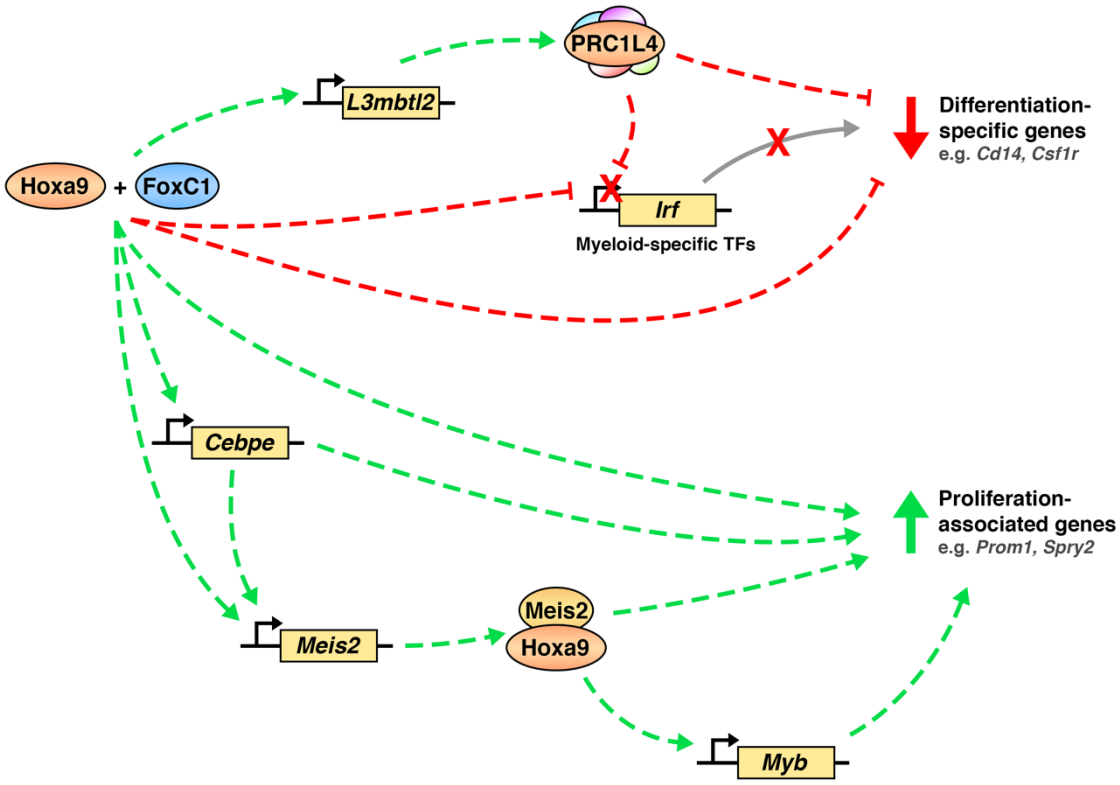


Figure 3:34 **A working model of the transcriptional network activated in Hoxa9+FoxC1 AMLs.**

Collectively, our analyses suggest a model in which Hoxa9 and FoxC1 collaborate to activate a transcriptional network defined by downstream activation of C/EBPε, Meis2 and Myb.

Hoxa9 and FoxC1 together with these upregulated TFs establish a specific program of gene expression promoting the survival and proliferation of Hoxa9+FoxC1 AML cells. We speculate that the widespread downregulation of the myeloid gene expression program is mediated in part by upregulation of the non-canonical Polycomb regulator L3mbtl2, which may directly repress differentiation-specific targets genes.

We identify *Meis2* as a putative target gene of FoxC1, and have thus far defined a collection of 19 *cis*-regulatory elements associated with *Meis2* activation spanning a ~250 kb chromatin domain. These sites include a probable 5' insulator element, buffering neighbouring genes from Hoxa9+FoxC1-dependent activation of *Meis2*, and a potential long-range enhancer element involving CTCF that might act upon the *Meis2* promoter by chromatin looping.

Prior to these studies, little was known about the *cis*-regulatory elements involved in the regulation of *Meis2*. As *Meis2* is a homeotic factor involved both in normal embryonic development and tumorigenesis, our efforts represent a meaningful contribution to the study of *Meis2* regulation that have implications beyond the study of this mouse model of AML. Significantly, despite upregulation of *Meis2* in Hoxa9+FoxC1 AMLs and previous reports of functional redundancy between Meis proteins, the gene regulation network activated in Hoxa9+FoxC1 cells are clearly distinct from that of Hoxa9+Meis1 cells, and these differences may arise directly from FoxC1 itself, or indirectly from other TFs deregulated as a result of FoxC1 expression.

Finally, integration of our analyses with those from primary FLT3-ITD human AML samples has demonstrated that the Hoxa9+FoxC1 AML gene expression signature shares some commonalities with these human primary AMLs, underlining the value of this mouse model as a relevant tool to gain insights into human disease.

3.23 Current limitations and future prospects

Despite the insights our analyses have generated thus far into the roles of Hoxa9 and FoxC1 in leukaemogenesis, we recognise that there are important limitations to the work as it currently stands.

Firstly, a major shortcoming of these studies so far arises from our reliance on the only publicly available FoxC1 ChIP-seq dataset, which was generated from an unrelated embryonic tissue. The fact that Hoxa9+FoxC1 AMLs and the tissue used for FoxC1 ChIP-seq are both distinguished by high *Meis2* expression[351] increases confidence in our analyses at this particular locus, but makes comparisons elsewhere challenging. To overcome this, we attempted to perform FoxC1 ChIP-seq analyses using chromatin prepared from a representative Hoxa9+FoxC1 AML sample, but were unsuccessful in

obtaining a meaningful signal enrichment (data not shown). Obtaining robust FoxC1 ChIP-seq data results was an ongoing challenge throughout these studies, and is the subject of further discussion in Chapter 6: Overall Discussion.

Secondly, despite a strong association between FoxC1 expression and the Hoxa9+FoxC1-specific upregulation of DHSs and proto-oncogenes including *Myb*, *Meis2* and others, the most reliable validation of target genes for a given TF such as FoxC1 is through further genetic perturbation experiments. To this end, in the latter months of this study we entered into a collaboration with Keith Humphries, who generously shared with us an immortalised murine leukaemia model generated from HSPCs transformed with the oncogene *Mn1*[387,388]. These cells not only feature robust expression of *Hoxa9* and *Foxc1,* but also express high levels of *Meis2*, which was modified using CRISPR to include an epitope tag together with a fluorescent mCherry reporter gene which is expressed in parallel with *Meis2* as a polycistronic transcript. These features allow reliable tracking of endogenous *Meis2* expression in live cells using FACS, while the epitope tag will greatly facilitate downstream analyses of *Meis2* in the absence of a reliable *Meis2* antibody. Consequently, these cells may represent a viable immortalised cell model in which to test some outstanding questions generated from our findings so far, including the validation of FoxC1-dependent target genes and *cis*-regulatory elements.

At the time of writing, we are generating lentiviral shRNA expression constructs targeting *Foxc1* to transduce into MN1-dependent AML cells to validate candidate FoxC1 targets, including *Meis2, Myb* and the MHC class genes. These data will provide critical insights into which genes are directly regulated by FoxC1 in the context of mouse AMLs, as well as potentially identify novel sensitivities common to human AML patients that could be targeted therapeutically. We also plan to directly validate putative Hoxa9+FoxC1 targets

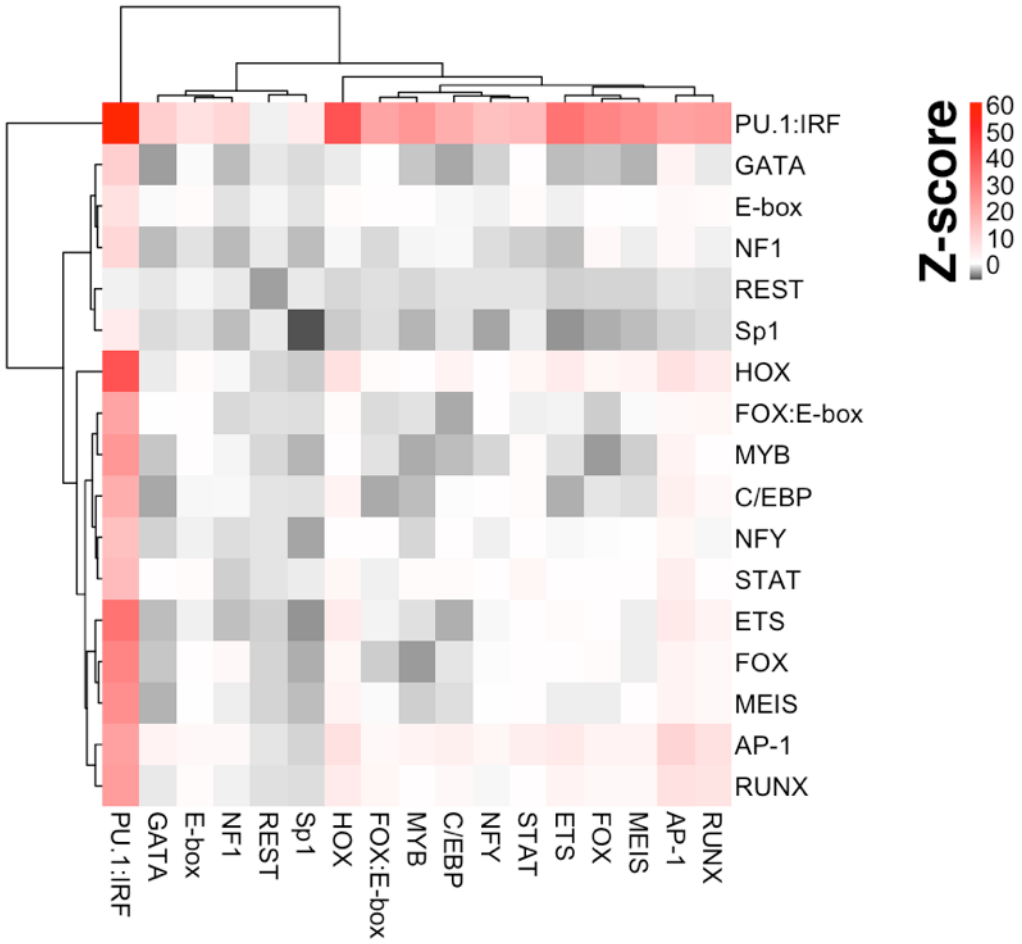
using similar shRNA knockdown approaches against downstream targets, coupled with phenotypic assays to examine changes in proliferation and differentiation.

Thirdly, several key questions remain as to how Hoxa9 and FoxC1 promote the widespread repression of genes associated with myeloid differentiation. One possibility is that *Cebpe* expression leads to downregulation of the myeloid TFs C/EBPα and C/EBPβ by feedback inhibition, suppressing the activation of normal myeloid differentiation pathways. We also identified a key component of the PRC1L4 complex as being substantially upregulated in Hoxa9+FoxC1 AMLs, but so far our observations are purely correlative and provide only indications of a potential mechanism. However, there are several approaches that could be taken to test this model. We could perform L3mbtl2 ChIP-seq in chromatin prepared from Hoxa9+FoxC1 AMLs to determine whether L3mtbl2 is localised to repressed loci and, if so, whether this is associated with an enrichment for the repressive H2AK119ub it deposits at target genes. By comparison of these data with those from Hoxa9+MMTV cells as a control, we could determine whether L3mtbl2 is a downstream mediator through which FoxC1 acts to promote the very striking repression of differentiation-specific genes we observed. These studies could be coupled with shRNA-mediated suppression of *L3mtbl2* in the immortalised Mn1-dependent AML cell model to directly assess the effect of L3mtbl2 depletion on H2AK119ub deposition, and/or the expression of these differentiation-specific genes.

Critically, our observations to date are based on bulk population-based analyses of leukaemic cells. As a result, insights into potential subclonal diversity within these samples are impossible to detect. The concept of subclonal heterogeneity in mouse models of AML was previously a topic of contention, but more recent studies suggest that transgenic mouse models of AML and B-ALL do indeed establish a clonal hierarchy as observed in primary human AMLs[389,390]. We consider that this is even more likely to

occur in the transplantation models of AML we used in this study, where donor cells co-transduced with Hoxa9 and FoxC1 or Meis1 are likely to be highly polyclonal, owing to variable viral transduction efficiencies, and are also probably subjected to clonal selection during engraftment in recipient mice. With the advent of reliable single-cell ATAC-seq and RNA-seq methods, a logical progression for this study would be to investigate the clonal composition of these cells, and the epigenetic and transcriptional signatures that underpin phenotypic differences between clones, if they exist.

Finally, the accessible chromatin landscape of Mn1-dependent AML cells could be profiled to compare their global similarity to the Hoxa9+FoxC1 AML model. If DHSs at the *Meis2* locus are identified that are common to the Hoxa9+FoxC1 model, the function of these *cis*-regulatory elements could be explored using classical enhancer reporter gene approaches, or by targeted perturbation of candidate enhancers/TF binding sites using genome-editing approaches such as CRISPR.



Appendix 3:1 **Motif bootstrapping analyses performed on 2270 Hoxa9+MMTV-specific DHSs reveals that these sites are overwhelmingly enriched for PU.1:IRF binding motifs.**

Initial motif bootstrapping analyses revealed that Hoxa9+MMTV-specific DHSs were dominated by an enrichment of PU.1:IRF binding sites, masking any other significant motif co-association patterns also present within these DHSs.

Appendix 3:2 **All significantly Hoxa9+FoxC1-deregulated genes (log2 FC >2.0 vs Hoxa9+MMTV)**

mRNA expression fold changes for all genes with Hoxa9+FoxC1-specific upregulation of DHSs and mRNA relative to both Hoxa9+MMTV and Hoxa9+Meis1

Gene name	log2 FC (Hoxa9+FoxC1 vs Hoxa9+MMTV)	log2 FC (Hoxa9+FoxC1 vs Hoxa9+Meis1)
Rbp4	7.29	7.29
Olfm4	5.79	5.37
Fmo2	5.34	5.32
Piezo2	4.73	4.73
Hdc	4.70	4.31
Grm4	4.53	4.56
Rbm24	4.53	4.58
Cyp2a5	4.50	4.61
Rgs5	4.45	4.47
Prom1	4.45	3.91
Lypd6b	4.19	4.55
Vcam1	4.16	3.78
Penk	4.16	4.65
Ffar2	4.09	4.89
Meis2	4.01	4.04
Tshr	3.99	4.72
Cnmd	3.98	4.10
Gatm	3.94	1.87
Olfml2b	3.94	4.58
Bmpr1a	3.90	4.58
Sdc2	3.86	3.97
Unc79	3.79	3.80
Glod5	3.77	3.77
Il20ra	3.71	3.66
H2afy2	3.54	3.85
Pgbd5	3.50	3.49
Smoc1	3.50	3.39
Plcxd3	3.47	3.47
Il1f9	3.35	3.80
Igf2bp2	3.35	3.47
Serpina11	3.32	3.78
Lrrc1	3.31	3.38
Gja10	3.28	3.35
Krt18	3.28	3.29
Sncaip	3.24	3.24
Ly75	3.16	3.97
Fkbp11	3.14	3.08
Rasgef1a	3.13	3.52
Slc5a11	3.12	2.94
Dhcr24	3.00	2.11
Dtwd1	2.94	2.38
Sox5	2.91	2.93
Ifi27	2.87	3.18
Adgrb1	2.82	3.10
Lin28a	2.79	3.13
Cmc2	2.74	2.41
Mark1	2.72	2.71
Slc40a1	2.70	2.38
Tnfrsf26	2.70	2.61
Ednra	2.61	2.28
Vegfa	2.51	2.90
Aff2	2.43	1.85
Cgnl1	2.42	1.30
Celsr1	2.40	2.37
Slco4c1	2.38	3.26
Lbp	2.34	1.43
Ffar4	2.34	2.27
Sult2a3	2.33	2.33
Slco3a1	2.25	1.52
Egln3	2.25	2.41
C130026I21Rik	2.24	2.41
Abhd5	2.23	2.26
Lgr4	2.19	1.85
Ada	2.15	1.52
Capn3	2.14	2.19
Bdnf	2.13	2.16
Creb3l3	2.12	1.59
Rida	2.09	2.03
Btbd11	2.09	1.70
Sun2	2.08	2.33
Sned1	2.08	1.97
L3mbtl2	2.06	1.90
Ttpa	2.04	2.12
Xrcc5	2.00	2.44

mRNA expression fold changes for all genes with Hoxa9+FoxC1-specific downregulation of DHSs and mRNA relative to both Hoxa9+MMTV and Hoxa9+Meis1

Gene name	log2 FC (Hoxa9+FoxC1 vs Hoxa9+MMTV)	log2 FC (Hoxa9+FoxC1 vs Hoxa9+Meis1)
Sept11	-2.43	-2.47
Ahnak	-6.32	-6.44
F13a1	-6.22	-6.61
Ptpro	-5.73	-5.73
Itga1	-5.64	-5.52
Lilrb4a	-5.46	-5.40
Lilr4b	-5.35	-5.33
Klf4	-5.16	-5.26
Lyz1	-5.11	-4.88
Emilin2	-5.11	-5.09
Met	-5.08	-4.79
Irf5	-5.08	-5.07
Capn2	-5.06	-5.28
Arhgef10l	-4.99	-5.29
Xdh	-4.97	-4.96
Hpse	-4.81	-4.38
Il10ra	-4.78	-4.43
Nlrp3	-4.77	-5.09
Vcan	-4.65	-5.72
Fam129b	-4.62	-4.63
Hip1	-4.57	-4.52
Abca6	-4.48	-5.49
Ctss	-4.47	-4.68
Chn2	-4.45	-4.48
Pld4	-4.44	-4.84
Ccr2	-4.42	-4.88
Dpp4	-4.41	-5.42
Clec4a3	-4.38	-4.51
Rubcnl	-4.32	-4.47
Rasgrf2	-4.31	-4.15
Apba1	-4.30	-4.30
Epha2	-4.28	-4.89
Cd163	-4.25	-5.03
Ext1	-4.23	-4.22
Abcb1b	-4.23	-3.69
Dusp3	-4.18	-4.84
Il6st	-4.11	-4.09
Plekhm3	-4.06	-4.23
Irf8	-4.04	-3.87
Angptl2	-4.02	-4.24
Crybg3	-4.01	-3.81
Trpm2	-3.96	-3.64
Mx1	-3.94	-3.78
Ddit4	-3.92	-4.13
Plxnb2	-3.91	-4.47
Rassf4	-3.89	-3.61
Dse	-3.85	-3.40
Zfyve9	-3.84	-3.09
Eif4e3	-3.80	-3.54
Cd48	-3.77	-3.90
Calhm2	-3.76	-3.74
Plekho1	-3.74	-3.82
Crip1	-3.73	-3.86
Ly86	-3.72	-4.04
Havcr2	-3.71	-3.66
Lpin1	-3.69	-3.62
Tns1	-3.69	-3.46
Tmem243	-3.63	-3.38
Ifitm6	-3.62	-4.85
Sulf2	-3.60	-3.03
Eps8	-3.59	-4.86
Frmd4b	-3.58	-4.07
Fam49a	-3.57	-3.27
Psap	-3.55	-3.30
Lyz2	-3.52	-3.54
P2ry1	-3.47	-3.52
Susd3	-3.46	-3.58
Dusp22	-3.44	-4.59
Tbxas1	-3.42	-3.37
Tcf4	-3.40	-3.30
Klf9	-3.40	-3.56
Ifi209	-3.36	-3.86
Cd2ap	-3.33	-3.80
Tns4	-3.33	-3.33

Appendix 3:2 (continued) **All remaining significantly Hoxa9+FoxC1-downregulated genes.**

mRNA expression fold changes for all genes with Hoxa9+FoxC1-specific downregulation of DHSs and mRNA relative to both Hoxa9+MMTV and Hoxa9+Meis1

Gene name	log2 FC (Hoxa9+FoxC1 vs Hoxa9+MMTV)	log2 FC (Hoxa9+FoxC1 vs Hoxa9+Meis1)
Nav1	-3.33	-2.29
Zcchc24	-3.32	-3.43
Sik1	-3.29	-3.16
Lpxn	-3.24	-3.78
Nfkbiz	-3.24	-2.96
Tpd52	-3.21	-3.40
Arap2	-3.20	-3.10
Bcl2	-3.18	-2.88
Itgb5	-3.18	-3.02
Hal	-3.10	-3.09
Slc37a3	-3.10	-3.01
Ccr1	-3.07	-4.04
Arhgap24	-3.07	-2.66
Mfsd6	-3.06	-2.85
Socs3	-3.01	-3.53
Inpp4a	-3.00	-2.55
Tsc22d3	-2.99	-3.38
Prkch	-2.99	-2.81
Mdfic	-2.97	-2.22
Slc8a1	-2.97	-2.58
Pag1	-2.96	-3.03
Btla	-2.96	-3.21
Hlcs	-2.96	-2.92
Stard3nl	-2.94	-2.88
Tmem176b	-2.88	-1.92
Arhgap26	-2.80	-2.61
Ccdc180	-2.77	-3.39
Rusc2	-2.72	-1.82
Rftn1	-2.71	-3.03
Tnfrsf1b	-2.71	-2.28
Trim3	-2.71	-2.81
Flnb	-2.69	-3.19
Zfp658	-2.69	-2.85
Kcna3	-2.69	-2.45
Atrnl1	-2.68	-2.37
Trim24	-2.67	-2.29
Eml6	-2.66	-2.07
Ahr	-2.66	-2.32
Elovl6	-2.65	-2.55
Zmiz1	-2.65	-2.45
Nrip1	-2.62	-2.42
Capg	-2.62	-2.62
Fos	-2.61	-3.26
Cyp4v3	-2.60	-1.89
Prkab2	-2.59	-2.29
Sema4d	-2.58	-2.22
Tulp4	-2.56	-2.38
Asprv1	-2.56	-2.68
Marcks	-2.54	-1.04
Zc3h12c	-2.54	-3.15
Map4k4	-2.53	-2.45
Zfp385a	-2.53	-2.52
Trim36	-2.52	-2.11
Zfp947	-2.49	-2.67
Slc29a3	-2.48	-2.09
Nsmaf	-2.46	-2.47
Kdm7a	-2.46	-2.46

mRNA expression fold changes for all genes with Hoxa9+FoxC1-specific downregulation of DHSs and mRNA relative to both Hoxa9+MMTV and Hoxa9+Meis1

Gene name	log2 FC (Hoxa9+FoxC1 vs Hoxa9+MMTV)	log2 FC (Hoxa9+FoxC1 vs Hoxa9+Meis1)
Asph	-2.45	-1.99
Ssbp2	-2.44	-2.61
Tifa	-2.40	-2.82
Sash1	-2.40	-2.45
Pparg	-2.40	-2.05
Gbp3	-2.40	-2.14
Unc119b	-2.39	-2.36
Iffo1	-2.39	-1.93
Gm9733	-2.38	-2.19
9030617O03Rik	-2.36	-2.79
Mitf	-2.36	-1.87
Phkg1	-2.36	-1.98
Mef2a	-2.35	-2.51
Cib2	-2.34	-2.66
Pdgfrb	-2.34	-2.12
Ncor2	-2.34	-1.72
Pcdh18	-2.32	-2.71
Cyp4f18	-2.30	-2.55
Pik3r5	-2.29	-2.83
Trio	-2.28	-2.18
Asap2	-2.28	-1.78
Prcp	-2.27	-2.47
Lgmn	-2.27	-2.52
Fabp5	-2.26	-2.84
Fyn	-2.26	-1.56
Plxdc1	-2.25	-1.89
Prokr1	-2.23	-1.85
Ank	-2.23	-2.15
Maml2	-2.22	-1.80
Tnfaip3	-2.21	-1.84
S1pr3	-2.20	-1.92
Jun	-2.20	-2.10
Prune2	-2.18	-1.33
Ube2h	-2.17	-2.06
Cd300a	-2.17	-2.18
Zyx	-2.17	-1.72
E030018B13Rik	-2.17	-2.02
Ica1l	-2.15	-2.47
Ctsl	-2.15	-2.84
Ly96	-2.13	-2.04
Gpr68	-2.12	-2.32
Bmt2	-2.12	-1.49
Al846148	-2.11	-1.81
Gltp	-2.11	-2.06
Stk32c	-2.10	-1.72
Rhobtb1	-2.10	-1.88
Klf13	-2.08	-2.36
Anxa4	-2.08	-2.07
Cped1	-2.06	-2.12
Stox2	-2.05	-1.12
Fam78b	-2.05	-2.68
Acsf2	-2.05	-1.64
Adgre5	-2.05	-2.36
Mkrn1	-2.05	-1.96
Zeb2	-2.02	-2.06
Prr5	-2.01	-1.80

Chapter 4: Molecular characterisation of FUJIOKA cells, an acute myeloid leukaemia cell line featuring high expression of *FOXC1*.

One of the aims of this study was to identify more appropriate cell models to study FoxC1 in the context of human AMLs. We determined that FUJIOKA human AML cells expressed comparatively high levels of *FOXC1* mRNA (Figure 3:1), therefore we sought to characterise epigenetic and transcriptional changes occurring as compared to other cells with low *FOXC1* expression. In this chapter I will investigate the chromatin, TF binding and gene expression profiles in FUJIOKA cells, and how these features may be related to the presence of FoxC1.

4.1 FUJIOKA cells feature a specific accessible chromatin signature enriched for RUNX, C/EBP and ETS binding motifs

To identify the complement of accessible chromatin regions present in FUJIOKA cells, we performed DNase-seq and compared these data to those generated from K562 cells[151], which feature comparatively modest levels of *FOXC1* expression. These analyses revealed the existence of 14,592 DHSs specific to FUJIOKA cells, 36,379 DHSs common to both FUJIOKA and K562 cells, and 27791 DHSs specific to K562 cells (Figure 4:1A).

Focusing on the 14,592 FUJIOKA-specific DHSs, HOMER *de novo* motif discovery analyses revealed a robust enrichment of ETS, C/EBP and RUNX binding motifs that were either less strongly-enriched or absent altogether in the 27,791 K562-specific sites (Figure 4:1C). FUJIOKA-specific DHSs also featured an enrichment of AP-1, E-box and FOX binding motifs, albeit at more modest levels. A search carried out within these DHS using the known FOX binding motif suggests that the FOX motif is present in FUJIOKA-specific and K562-specific DHSs (Figure 4:1B).

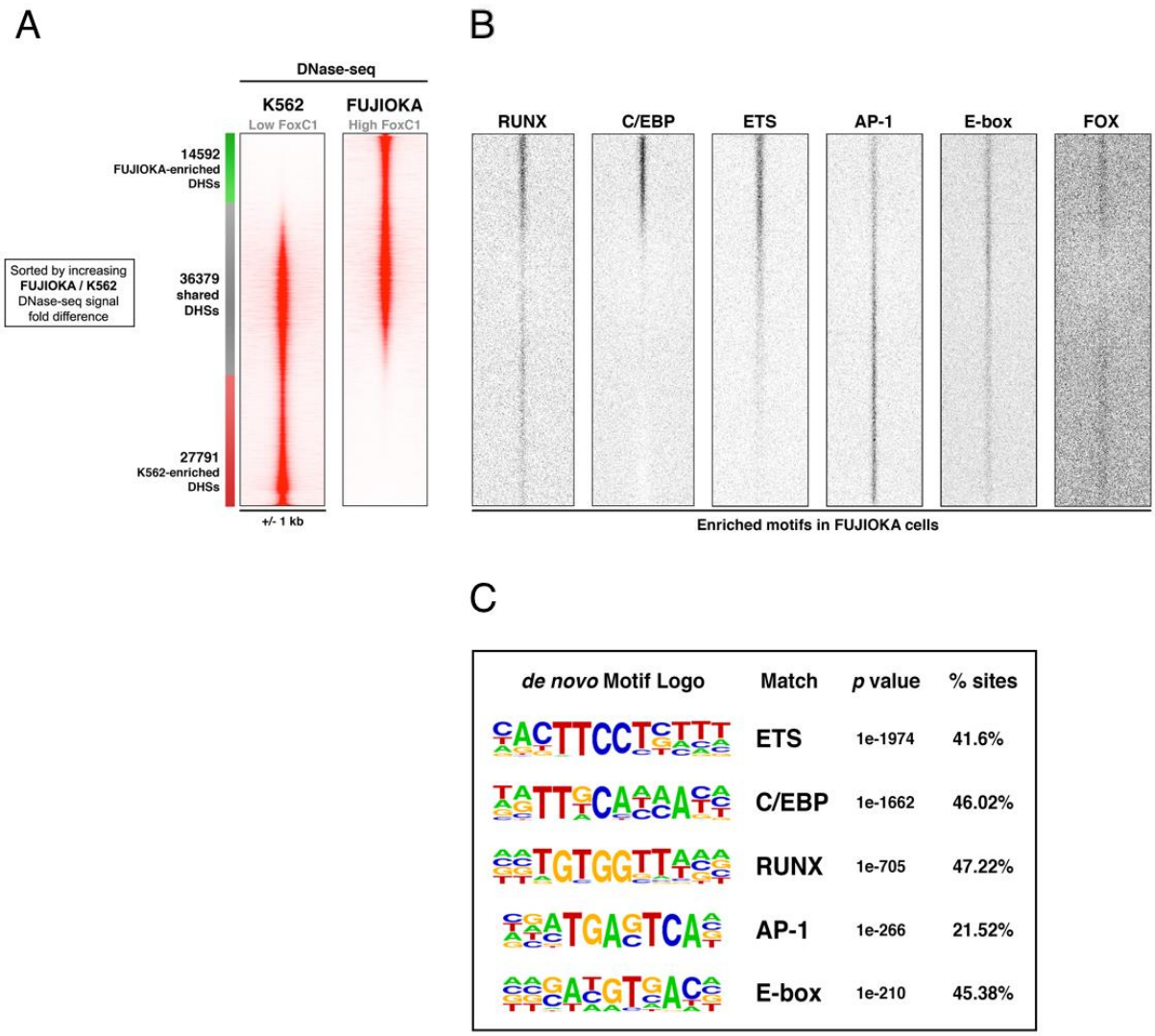


Figure 4:1 **FUJIOKA cells feature a specific chromatin signature differentially enriched for TF binding motifs.**

A: Profiles of DNase-seq signal for FUJIOKA cells and K562 cells, presented as the union of all peaks in both cell lines ranked in order of increasing FUJIOKA / K562 DNase-seq signal fold-difference. Profiles are displayed as 2 kb windows, centred on the summits of all DHSs. DHSs exhibiting greater than a +/- 4 fold change were classified as differential DHSs.

B: HOMER motif density plots for various TF binding sites enriched in FUJIOKA-specific DHSs. At far right is the FOX binding motif, which is not significantly enriched, but abundantly distributed in FUJIOKA-specific DHSs.

C: HOMER *de novo* discovery results for TF motifs significantly enriched in FUJIOKA-specific DHSs.

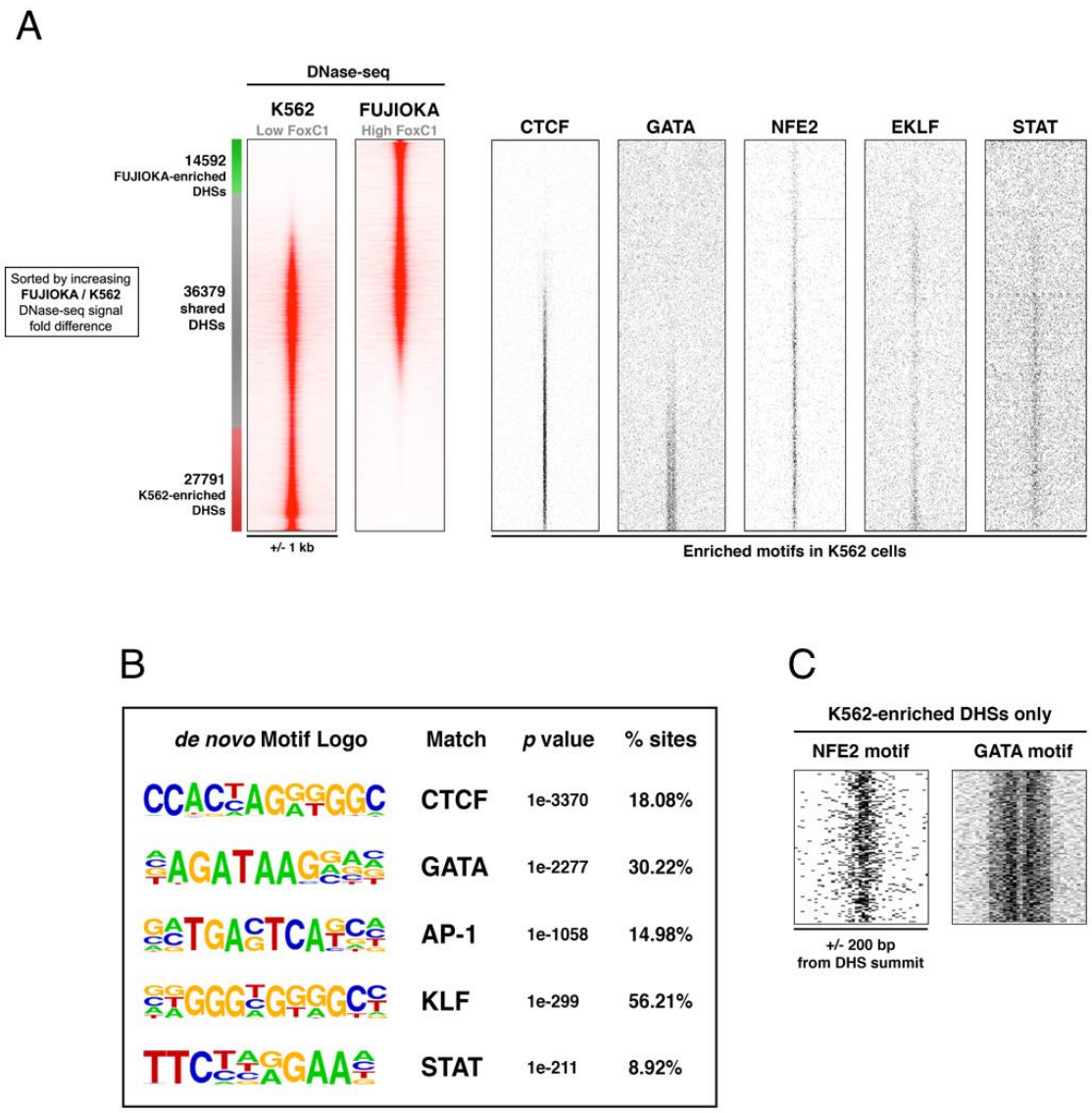


Figure 4:2 **K562-specific DHSs are enriched for distinct TF binding sites to FUJIOKA DHSs.**

A: Profiles of DNase-seq signal for FUJIOKA cells and K562 cells, with HOMER motif density plots for various TF binding sites enriched in K562-specific DHSs plotted against the same co-ordinates.

B: HOMER *de novo* discovery results for TF motifs significantly enriched in K562-specific DHSs.

C: Detail of NFE2 and GATA motif enrichment in K562-specific DHSs, centred on the DHS summits as in **A** but across a narrower window of +/- 200 bp from the summit.

4.2 DHSs absent in FUJIOKA cells but activated in K562 cells are enriched for distinct classes of binding motifs

Next, we sought to identify specific features of the K562 cistrome that were absent in FUJIOKA cells by performing similar motif analyses on the group of 27,791 K562-specific DHSs (Figure 4:2B). At these DHSs we identified a specific enrichment of GATA, KLF, NFE2 and STAT binding motifs. NFE2 motif enrichment was not revealed by *de novo* motif discovery, but an AP-1 motif which could be recognised by NFE2 was enriched in the *de novo* analyses. These DHSs were also enriched for CTCF binding motifs together with the group of DHSs common to both K562 and FUJIOKA cells. Finally, a search within these DHSs using the known NFE2 binding motif (Appendix 4:1) revealed a preferential enrichment of NFE2 sites closer to the summits of K562-specific DHSs, while GATA sites were most strongly enriched either side of the summits of these DHSs (Figure 4:2C).

4.3 DHSs activated in FUJIOKA cells are distinguished by occupancy of RUNX1, C/EBPα and FoxC1

To investigate which TFs might play important roles in the control of the *cis*-regulatory elements specifically present in FUJIOKA cells, we performed ChIP-seq to profile the occupancy patterns of FoxC1, RUNX1 and C/EBPα, which all had enriched binding motifs in FUJIOKA-specific DHSs (Figure 4:1).

These analyses revealed robust occupancy of both RUNX1 and C/EBPα at FUJIOKA-specific DHSs, with the strongest binding occurring at DHS summits, in concordance with motif enrichment patterns (Figure 4:3B). RUNX1 and C/EBPα binding was also observed at more modest levels in the common group of DHSs shared by FUJIOKA and K562 cells. Similar to RUNX1 and C/EBPα, the binding of FoxC1 at FUJIOKA-specific DHSs occurred most strongly towards the summits of these DHSs. However, FoxC1 binding in the shared group of DHSs occurred predominantly at regions flanking the hypersensitive

sites, in stark contrast to RUNX1 and C/EBPα. These FoxC1-bound sites were not enriched for any motifs, although the hypersensitive sites they flanked did feature an overrepresentation of an unknown 5'-CGCCC(T/C)(T/C)T-3' motif.

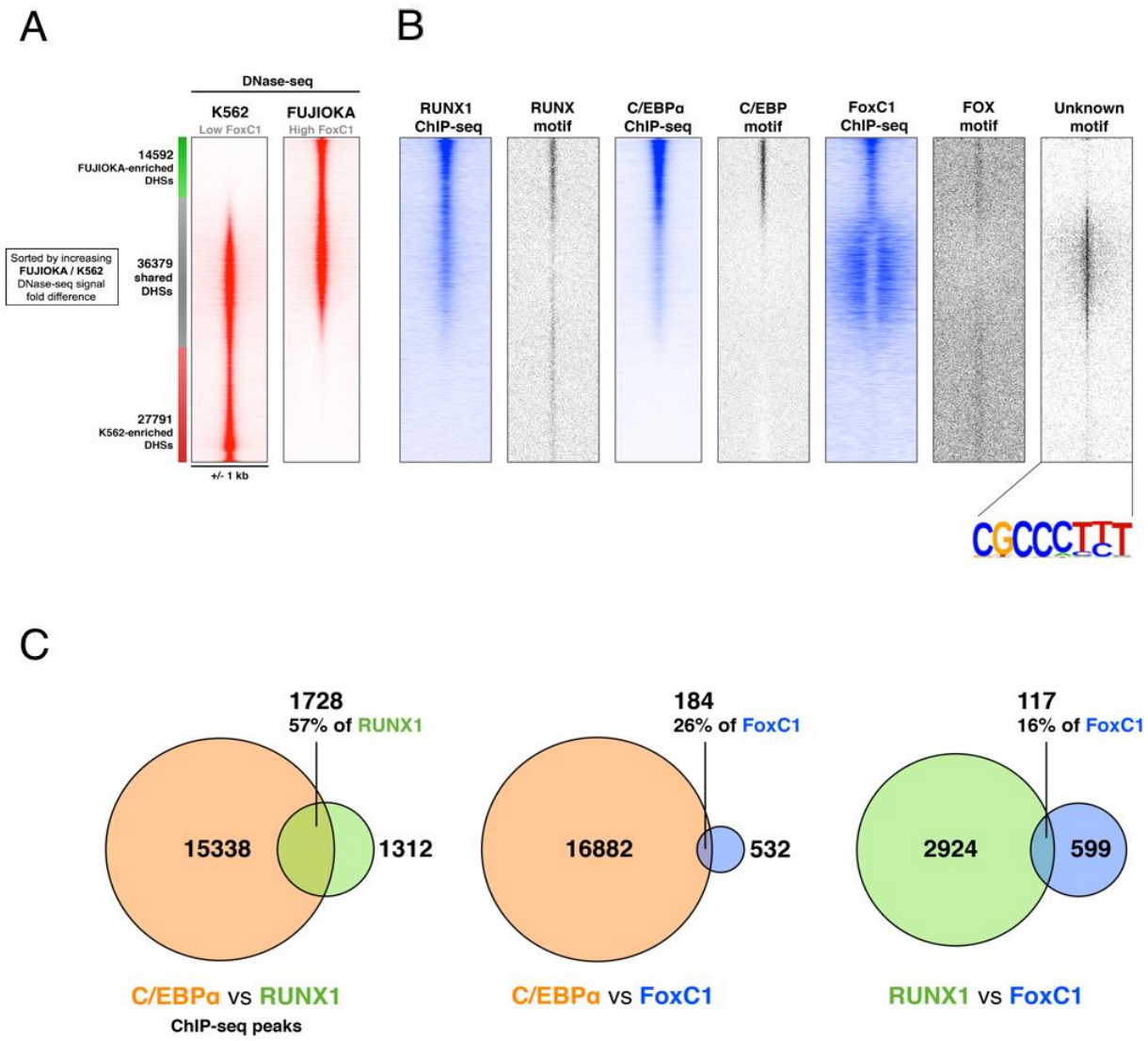


Figure 4:3 **FUJIOKA-specific DHSs are occupied by RUNX1, C/EBPα and FoxC1.**

A: Profiles of DNase-seq signal for FUJIOKA cells and K562 cells, presented as in Figures 4:1 and 4:2.

B: HOMER known motif density and TF ChIP-seq signal intensity, plotted on the same axis as **A**.

C: Venn diagram analyses showing pairwise comparisons of the highest-confidence peaks* in each of the three TF ChIP-seq datasets to identify common and discretely occupied loci for each pair of TFs.

*for this analysis, each ChIP-seq dataset was filtered to remove repeat elements and artefactual sequences from the ENCODE hg19 blacklist. A peak was considered high confidence if the number of uniquely-mapped reads exceeded the following thresholds: C/EBPα = **44**; RUNX1 = **7**; FoxC1 = **8**.

To investigate whether binding for the above TFs occurred within close proximity of each other, we performed an overlap of the highest-confidence ChIP-seq peaks for each dataset (Figure 4:3C). Indeed, 57% of RUNX1 binding events overlapped with C/EBPα binding, and 26% of sites bound by FoxC1 also overlapped with C/EBPα binding. The overlap between RUNX1 and FoxC1 binding was more modest, with around 16% of FoxC1-bound regions also being occupied by RUNX1 in these datasets.

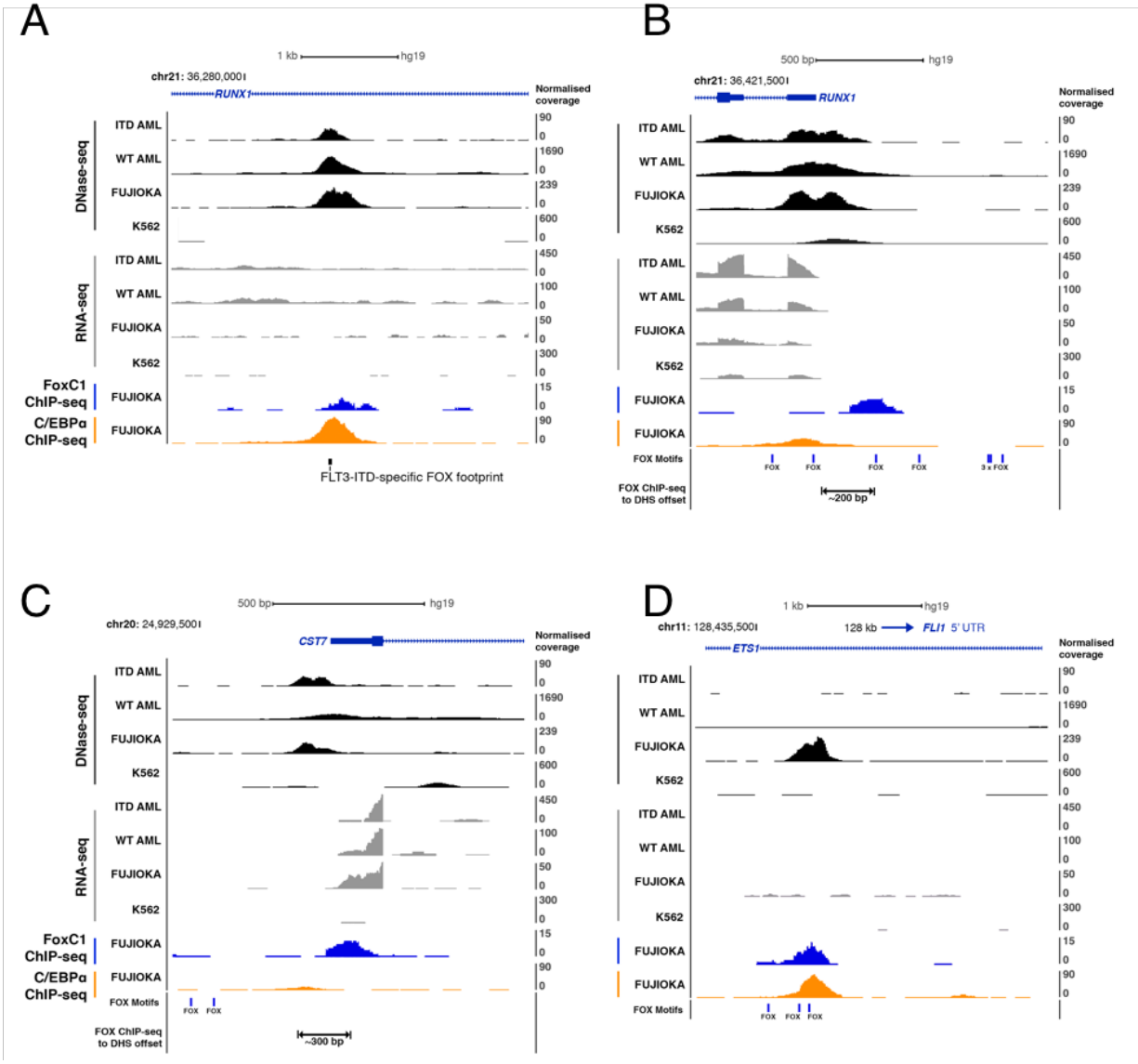


Figure 4:4 **A subset of DHSs activated in FUJIOKAs are occupied by C/EBPα and FoxC1.**

(Continued on next page)

(continued from previous page)

UCSC genome browser views of loci activated in FUJIOKA cells that also feature binding of FoxC1 and C/EBPα, as shown by DNAse-seq, RNA-seq and ChIP-seq analyses. Views are shown of two DHSs at the *RUNX1* locus (**A** and **B**), as well as others at the *CST7* (**C**) and *FLI1* (**D**) loci.

Samples are shown alongside a representative FLT3-ITD/NPM1c+ AML patient (ITD AML) and t(8;21) AML patient with wild type FLT3 (WT AML). These samples are named ITD/NPM1-2 and t(8;21)-4 in a manuscript currently being prepared by this laboratory.

Manual inspection of individual loci enriched for FoxC1 binding in FUJIOKA cells confirmed co-association with C/EBPα binding for a subset of sites which were also DHSs specific to primary FLT3-ITD AMLs (Figure 4:4). These included two DHSs activated at the *RUNX1* locus as well as two DHSs associated with the genes *CST7* and *FLI1*, all of which were found to be upregulated in FUJIOKA cells relative to K562 cells (Figure 4:7). Another FoxC1 binding peak at the intragenic *RUNX1* DHS occurred at a site previously determined to be specifically occupied in primary FLT3-ITD AML cells using high-resolution DNase I digital footprinting analyses[271]. Finally, as found elsewhere in FUJIOKA cells (Figure 4:3) and our *in vivo* experiments (Figure 3:31), we observed evidence of offset FoxC1 binding with respect to DHSs at the *RUNX1* and *CST7* loci. In these examples, the offset distance between maximal FoxC1 ChIP-seq signal and the DHS summit was ~200 and ~300 bp, respectively.

4.4 Regions bound by FoxC1 in FUJIOKA cells correlate with specific DHSs and TF occupancy patterns in primary FLT3-ITD AML samples

FOXC1 is upregulated in a high proportion of primary FLT3-ITD AML samples[271,300], which prompted us to investigate whether regions bound by FoxC1 in FUJIOKA cells correlated with FLT3-ITD-specific chromatin accessibility patterns from a representative primary sample, sample ITD-1 from Cauchy *et al.* (2015), as compared to normal $CD34^+$ PBSC cells (Figure 4:5). Indeed, we found that FoxC1 binding in FUJIOKA cells was preferentially enriched at regions corresponding to upregulated DHSs in primary FLT3-ITD AML samples. Furthermore, these regions were also distinguished by an enrichment

of FOX:E footprints from high-resolution DNase I digital footprinting analyses of the same primary FLT3-ITD AML sample.

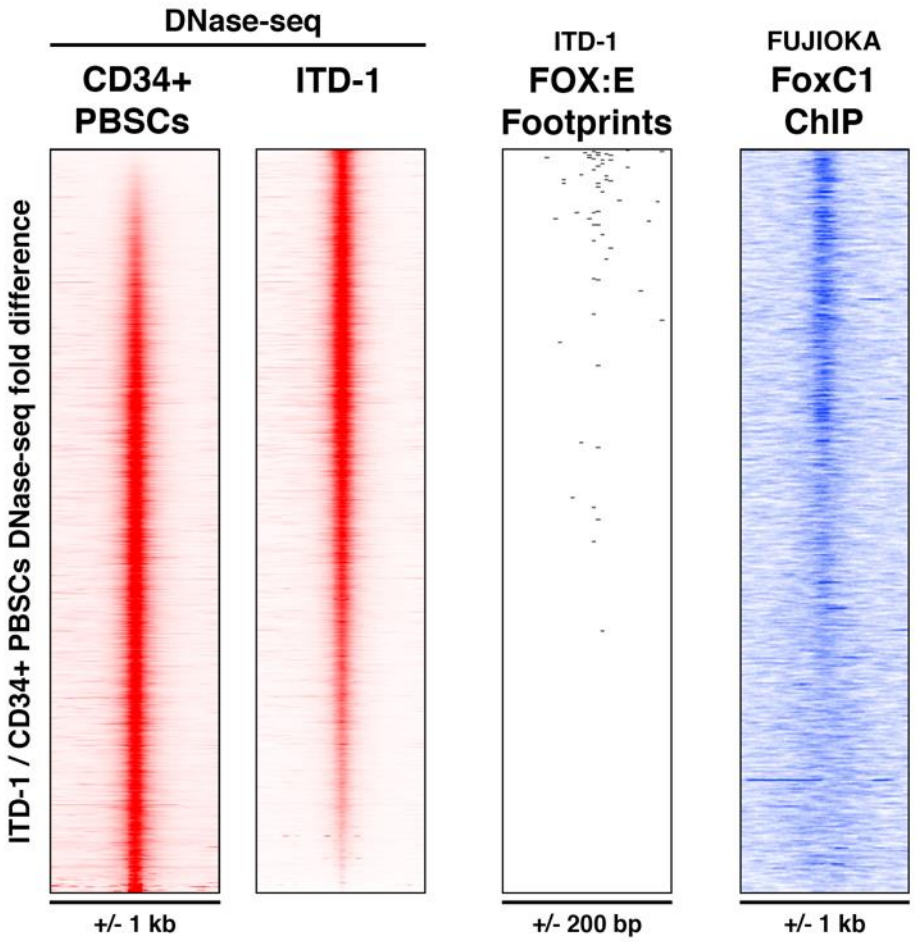


Figure 4:5 **FoxC1 binding events in FUJIOKA cells correlate with DHSs and FOX:E footprints in a representative FLT3-ITD AML patient.**

(Left) Analyses of DNase-seq data from normal human CD34+ PBSCs and a representative primary FLT3-ITD AML sample (ITD-1) from Cauchy *et al.* (2015). DNase-seq profiles are plotted in order of increasing fold difference in DNase-seq signal between ITD-1 and $CD34^{+}$ PBSCs.

(Right centre) FOX:E and FOX DNase-seq digital footprints identified in ITD-1, plotted against the same DHS coordinates as the DNase-seq data. For clearer visualisation of footprint positions within DHSs, these are plotted on a smaller window of +/- 200 bp relative to the same DHS summits as the DNase-seq data.

(Far right) FoxC1 ChIP-seq from FUJIOKA cells presented on the same genomic co-ordinates as the DNase-seq and FOX:E footprinting data.

4.5 Patterns of chromatin accessibility are associated with changes in gene expression between FUJIOKA and K562 cells

We sought to assess whether the differences in chromatin features identified between FUJIOKA and K562 cells were associated with specific changes in the expression of mRNAs by performing transcriptomic analyses of RNA-seq data.

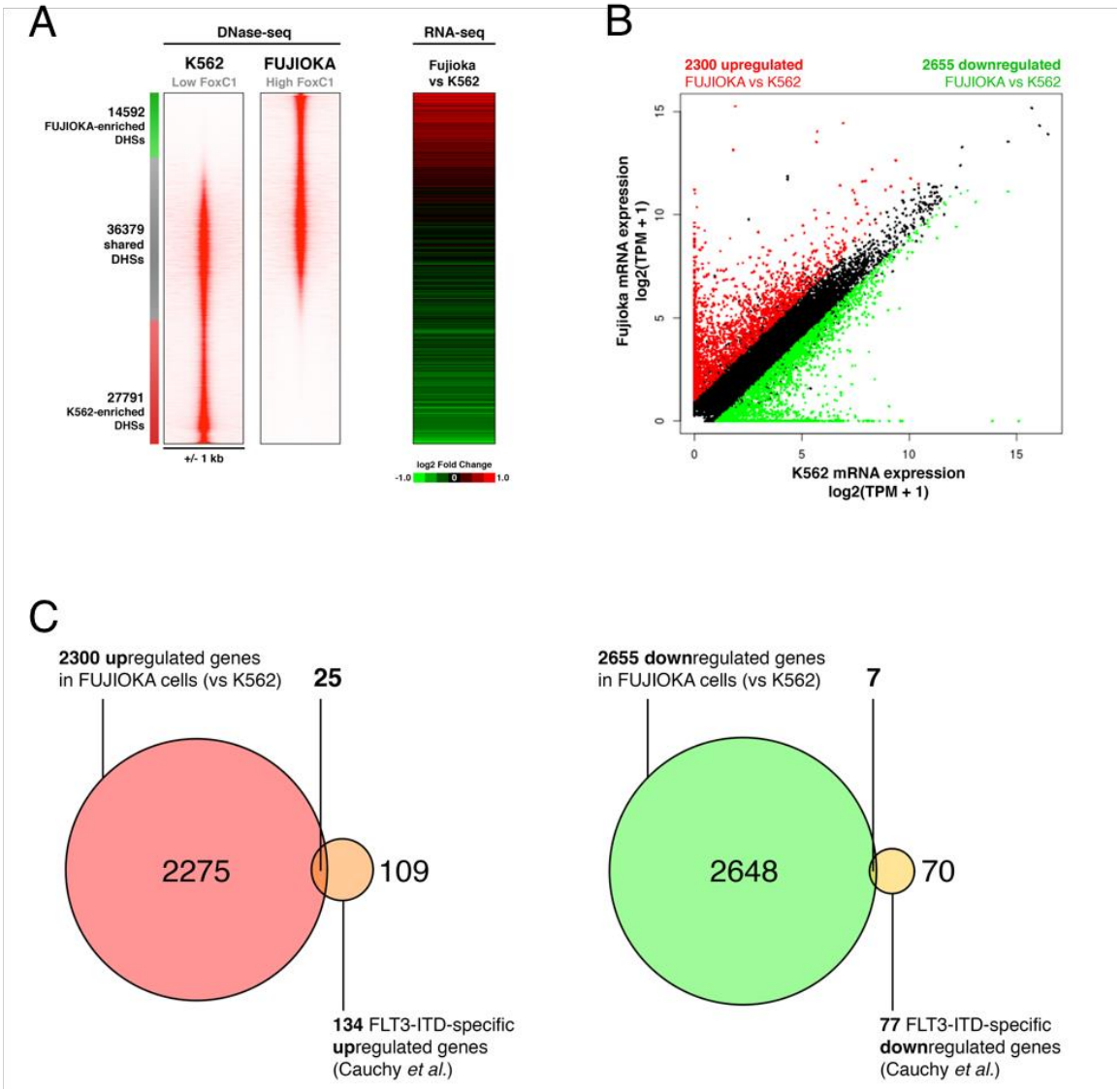


Figure 4:6 **Figure 4:6: DHSs specific to FUJIOKA and K562 cells correlate with changes with gene expression.**

A: (Right) Gene expression changes between FUJIOKA and K562 RNA-seq data, presented as the nearest genes to DHSs ranked by FUJIOKA / K562 DNase-seq fold difference (left).

(continued on next page)

(continued from previous page)

B: XY scatter plot of differentially expressed genes between FUJIOKA and K562 RNA-seq data. Significantly and substantially* upregulated genes in FUJIOKA cells are coloured in red, while similarly downregulated genes are coloured in green.

C: Venn diagram analyses showing pairwise comparisons of differentially-expressed genes in FUJIOKA cells with FLT3-ITD-specific genes identified in Cauchy *et al.* (2015).

*Genes were considered significantly and substantially changing if the expression values exceeded the following threshold: log(2) FC > 1.0, adJ $p < 0.05$.

Indeed, we found that changes in chromatin accessibility between the two cell lines correlated with broad shifts in gene expression (Figure 4:6A). The FUJIOKA-specific DHSs were associated with an upregulation of mRNAs relative to K562 cells. By contrast, DHSs specific to K562 cells were associated with genes downregulated in FUJIOKA cells. Furthermore, a pairwise comparison of duplicate RNA-seq datasets for both cell lines (Figure 4:6B) revealed the significant upregulation of 2300 genes and downregulation of 2655 genes in FUJIOKA cells as compared to K562 cells.

Strongly upregulated genes in FUJIOKA cells included *FOXC1* together with *HOXA9*, similar to observations made in primary AML samples (Figure 4:7). In concordance with an enrichment of ETS sites in FUJIOKA-specific DHSs, we observed an upregulation of the ETS family genes *FLI1* and *SPI1*. Furthermore, we identified FUJIOKA-specific upregulation of *CEBPA* and *RUNX1*, consistent with an enrichment of binding motif and TF occupancy patterns within FUJIOKA-upregulated DHSs (Figure 4:3). Conversely, FUJIOKA RNA-seq analyses revealed broad downregulation of a number of transcription factors recognising motifs enriched in K562-specific DHSs, including GATA, KLF, AP-1 and STAT family TF genes. *NFE2* was also downregulated in FUJIOKA cells, but less substantially than these other TF genes.

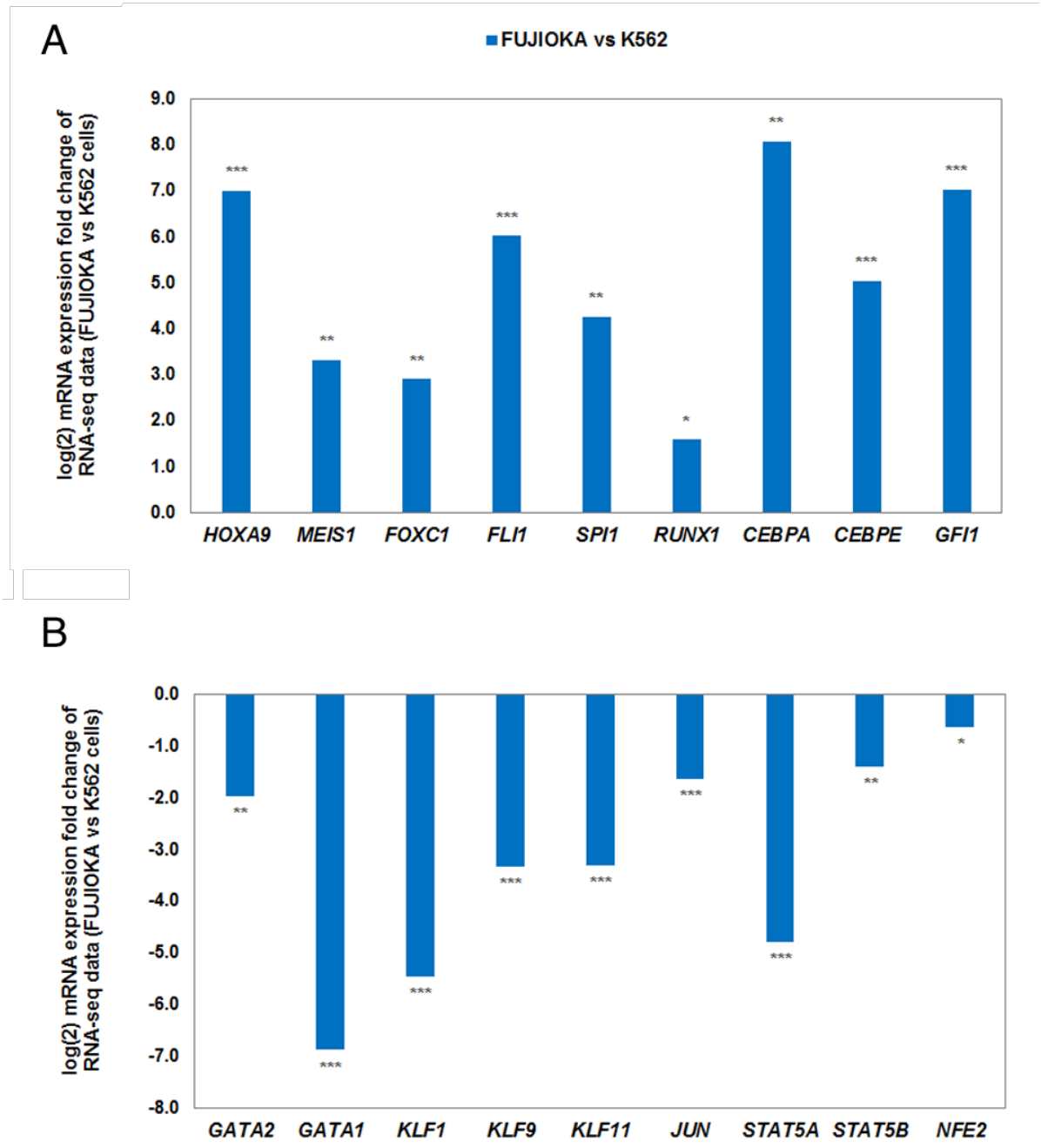


Figure 4:7 **FUJIOKA cells feature upregulation of *HOXA9* and *FOXC1*, together with several TF genes that recognise enriched motifs in FUJIOKA-specific DHSs.**

mRNA expression fold changes for differentially-expressed genes in FUJIOKA cells as compared to K562 cells, calculated from average normalised RNA-seq expression values across two replicates for each cell line.

A: Upregulation of TF genes that may recognise motifs enriched in FUJIOKA-specific DHSs.

B: Downregulation of TF genes that may recognise motifs enriched in K562-specific DHSs.

* = adj. P <0.05; ** = adj. P <0.01; *** = adj. P <0.001

Because another aim of this study was to identify alternative cell models to study *FOXC1* in the context of FLT3-ITD, we inspected the extent to which differentially regulated genes in FUJIOKA cells overlapped with the FLT3-ITD-specific gene expression signature previously identified in Cauchy *et al.* (2015) (Figure 4:6C). These analyses revealed a modest overlap between the two datasets, with 25 upregulated genes and seven downregulated genes present in both FUJIOKA cells and the FLT3-ITD gene list (Table 4:1).

Table 4:1 **Genes differentially regulated in FUJIOKA cells (vs K562 cells) are also deregulated in primary FLT3-ITD AML samples.**

mRNA expression fold changes for genes in FUJIOKA cells (vs K562 cells) commonly upregulated (≥ 2-fold, adj *P* <0.05) in the Cauchy *et al.* FLT3-ITD-specific gene list.

Gene name	log2 FC (FUJIOKA vs K562)	adj. *P*
CTSG	11.20	1.18E-06
IGFBP2	8.77	3.48E-06
CST7	8.04	2.64E-05
AZU1	7.93	3.21E-04
PRTN3	7.31	3.04E-04
SPNS3	5.65	5.09E-04
IRX3	5.43	3.20E-05
CCL5	5.39	7.11E-06
ELANE	4.85	2.26E-03
CLU	4.63	5.15E-05
MS4A3	4.52	2.47E-04
BIK	3.84	5.34E-04
CD99	3.61	9.91E-05
SIGLEC12	3.61	3.41E-05
MDFI	3.23	1.77E-03
FOXC1	2.91	1.05E-03
LTC4S	2.57	2.05E-02
MAGED1	2.39	1.85E-03
PBX3	1.84	9.00E-04
CRNDE	1.74	9.51E-03
SPRY2	1.73	5.60E-04
RASGRP3	1.46	5.03E-03
CACNB4	1.39	2.04E-02
RXFP1	1.15	1.07E-02
PARD6A	1.02	8.69E-03

mRNA expression fold changes for genes in FUJIOKA cells (vs K562 cells) commonly downregulated (≥ 2-fold, adj. *P* <0.05) in the Cauchy *et al.* FLT3-ITD-specific gene list.

Gene name	log2 FC (FUJIOKA vs K562)	adj. *P*
APOC1	-5.54	6.94E-05
HDGFRP3	-4.36	1.35E-05
JUP	-3.70	7.95E-04
TNFRSF10B	-2.52	3.75E-04
BLVRA	-1.88	9.27E-04
TRPS1	-1.77	2.13E-03
FNIP1	-1.09	1.56E-02

4.6 Preliminary data suggest that shRNA-mediated knockdown of FoxC1 disrupts proliferation of FUJIOKA cells

Previous studies for several leukaemia cell lines demonstrated that knockdown of endogenous levels of *FOXC1* expression disrupted proliferation and induced morphological differentiation[300], consistent with overexpression studies in normal human HSPCs and the Hoxa9+FoxC1 mouse model of AML, which led to enhanced cell proliferation and a block to normal differentiation. However, these studies did not yet extend to the FUJIOKA cell line, which also express comparatively high levels of *FOXC1* mRNA for a leukaemia cell line (Figure 3:1). Critically, these analyses did not explore how *FOXC1* suppression impacted the chromatin landscape and gene expression profile in any leukaemia cell line to date. To address this, we cloned lentiviral *FOXC1* shRNA vectors to introduce into FUJIOKA cells for *FOXC1* knockdown.

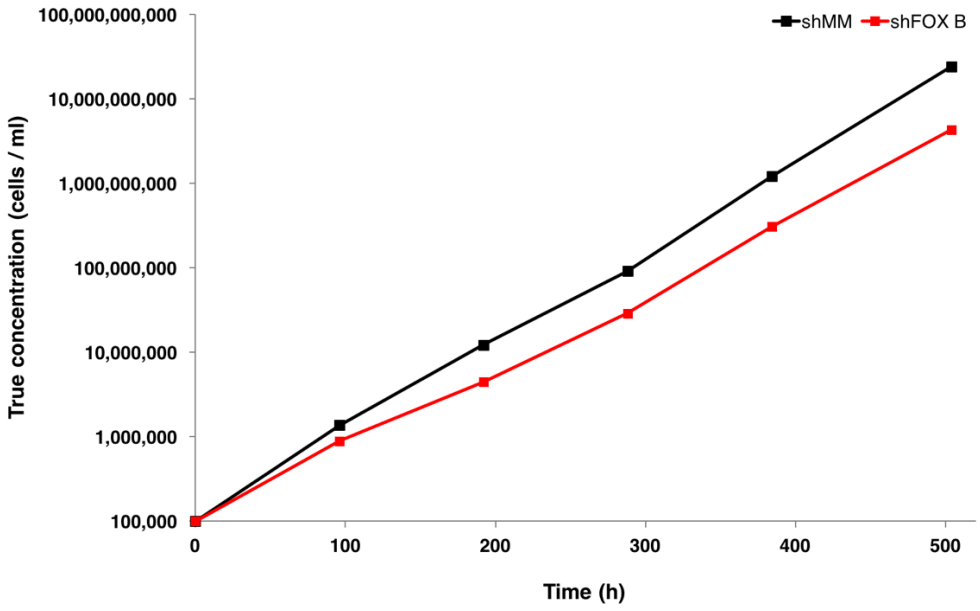


Figure 4:8 **Preliminary data suggests that transduction of FUJIOKA cells with an shRNA targeting *FOXC1* impairs cell growth.**

Cell growth curves for FUJIOKA cells transduced with lentiviral expression vectors for shRNAs targeting *FOXC1* (shFOX B), together with a scrambled non-targeting control (shMM). Cell numbers are expressed as the 'true' concentration, i.e. based on retention of all cells at each passage. The preliminary data presented here is from a single replicate for each shRNA transduction.

At the time of writing, owing to unforeseen delays in generating shRNA expression vectors, we were unable to complete these experiments in time for inclusion here. However, an initial pilot experiment (of just one replicate thus far) suggests that introduction of one *FOXC1*-targeting shRNA (shFOX B) into FUJIOKA cells caused a delay in cell growth as compared to a mismatch shRNA control. (Figure 4:8). We are now profiling the mRNA and protein expression of *FOXC1* in these samples and have begun replicate experiments both with the constructs described, as well as with vectors expressing two further shRNA sequences targeting *FOXC1* mRNA, to account for potential off-target effects.

4.7 CRISPR deletion of the *FOXC1* gene coding region

To investigate the epigenetic and transcriptional consequences arising from complete ablation of *FOXC1* expression in AML cells, we used a CRISPR/Cas9 genome-editing strategy to introduce a targeted deletion of the *FOXC1* locus in FUJIOKA cells (Figure 4:9). A pair of single guide RNAs (sgRNAs) targeting a region immediately downstream of the *FOXC1* start codon and another pair targeting the stop codon were designed, with the aim of deleting almost all of the coding region of *FOXC1*. Two independent pairs of sgRNA sequences were used for each site targeted to mitigate against potential off-target effects (full experimental details can be found in Materials & Methods 2.11.1).

From these experiments, 2 viable clones of ~750 transfected single cells were found to carry heterozygous deletion for *FOXC1*, as determined by PCR analyses (Figure 4:9B). qPCR analyses of *FOXC1* mRNA expression revealed an increase in *FOXC1* mRNA in CRISPRΔ1 clone 4 relative to wild-type, whereas the other clone (CRISPRΔ2 clone 1) featured a modest but significant ($p < 0.001$) decrease in *FOXC1* mRNA as compared to wild-type FUJIOKA cells. Finally, we note that thus far, we were unable to obtain viable clones carrying a homozygous *FOXC1* deletion (data not shown).

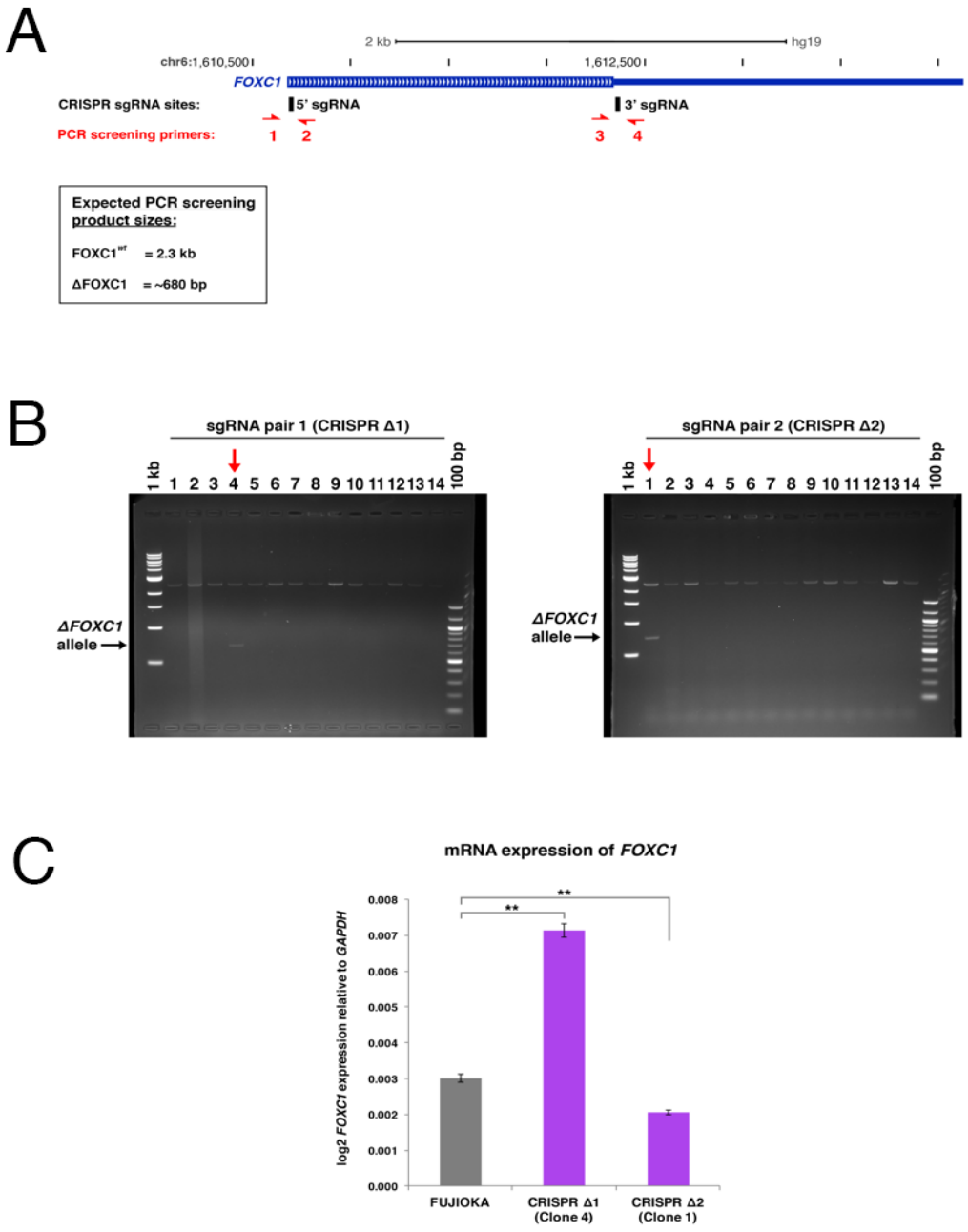


Figure 4:9 **CRISPR deletion of the *FOXC1* gene in FUJIOKA cells.**

A: CRISPR deletion strategy for the coding region of *FOXC1*. Indicated underneath the *FOXC1* gene map are the regions targeted by pairs of sgRNAs (black) and by sets PCR primers (red) to validate deletions.

B: Agarose gel electrophoresis of PCR products (using screening primers 1 and 4) reveals heterozygous *FOXC1* deletions (*ΔFOXC1*) in one clone for each independent pair of sgRNAs tested. Positive clones of FUJIOKA cells are indicated with red arrows.

C: qRT-PCR analyses of *FOXC1* mRNA expression in normal FUJIOKA cells as compared to the two heterozygous *FOXC1*$^{wt/\Delta}$ clones which grew. Error bars plotted show the standard error of technical qRT-PCR replicates. ** = p <0.001, as measured by paired student *t*-test.

Chapter 4: Discussion

4.8 Transcription factor motif enrichment patterns in FUJIOKA and K562-specific DHSs broadly reflect differences in lineage bias

Through DNase-seq analyses, we have identified a specific accessible chromatin signature present in FUJIOKA cells but not K562 cells (Figure 4:1). DHSs present in FUJIOKA cells but not K562 cells were distinguished predominantly by an enrichment of ETS, RUNX and C/EBP binding sites, but also a modest enrichment of FOX motifs. An overrepresentation of ETS and C/EBP motifs, which are recognised by the myeloid-associated TFs PU.1 and C/EBPα, is consistent with the monocytoid phenotype of FUJIOKA cells, which were established from a primary acute monoblastic leukaemia sample[391]. Indeed, analyses of RNA-seq data from FUJIOKA cells confirmed that *SPI1* (PU.1) and *CEBPA*, both of which encode myeloid lineage regulators, were upregulated as compared to K562 cells (Figure 4:7), together with the granulopoietic TF gene *CEBPE*[366]. However, we also noted substantial upregulation of the proto-oncogene *FLI1*, which may bind ETS motifs preferentially enriched in FUJIOKA-specific DHSs[392,393].

By contrast, the global DNase-seq profile of K562 cells was distinguished from FUJIOKA cells by an enrichment of binding sites for GATA, KLF and NFE2 factors (Figure 4:2). K562 cells are a myelogenous leukaemia cell line derived from a primary blast crisis CML sample, and have the potential to differentiate into monocytes/macrophages and granulocytes[394], but they also bear substantial erythroid characteristics, including expression of the erythrocyte marker glycophorin[395]. Thus, it is not surprising that the DHSs specific to K562 cells were enriched for binding sites which are recognised by TFs involved in erythroid lineage commitment, including GATA1 and KLF1 (EKLF), both of which were upregulated in K562 cells as compared to FUJIOKA cells (Figure 4:7).

Our observations that NFE2 binding sites were enriched at the summits of K562-specific DHSs (Figure 4:2) in contrast to GATA sites, which tended to flank DHS summits, recapitulates earlier locus-specific analyses on a genome-wide scale. For example, previous studies of the human beta-globin LCR demonstrated that activation of DHS '2', the only site in the LCR to act as a transcriptional enhancer, was dependent upon prior binding of NFE2 which, via recruitment of ATP-dependent chromatin remodelling enzymes, caused localised opening of chromatin for GATA1 and EKLF to subsequently gain access to target motifs flanking this pre-bound NFE2 site[396].

4.9 The FUJIOKA-specific accessible chromatin signature is distinguished by co-localised binding of C/EBPα and FoxC1

Because we observed an enrichment of RUNX, C/EBP and FOX binding sites in DHSs specific to FUJIOKA cells (Figure 4:1), we performed ChIP-seq analyses for three abundantly-expressed TFs recognising these sites, including FoxC1. Firstly, these analyses confirmed that DHSs specific to FUJIOKA cells and shared with K562 cells were occupied by RUNX1. Occupancy of sites present in these phenotypically dissimilar haematopoietic cell types is consistent with extensively documented roles of RUNX1 as a master regulator of haematopoiesis, where its expression is essential not only for the emergence of primitive blood precusors during embryogenesis, but also throughout adult haematopoiesis including differentiation into erythroid, myeloid and lymphoid lineages[210,397,398].

Secondly, an enrichment of C/EBPα binding in FUJIOKA-specific DHSs is – at a superficial level – reflective of the myeloid lineage bias and higher relative expression levels of *CEBPA* mRNA in FUJIOKA cells as compared to K562 cells[230,236]. Similarly, robust FoxC1 binding within FUJIOKA-specific DHSs is concordant with substantially higher levels of *FOXC1* expression in these cells as compared to K562 cells. Critically,

we also observed that *FOXC1* upregulation in FUJIOKA cells was paralleled by a robust increase in *HOXA9* mRNA expression, which is strongly downregulated in differentiating haematopoietic cells[329]. These two observations reflect other studies – including those discussed in Chapter 3 – that highlight a phenotypically evident but mechanistically unclear co-operativity between HOXA9 and FoxC1 in AML[300].

Because other FOX proteins are known to interact with a myriad of other TFs to establish highly context-specific patterns of gene regulation[399,400], we explored the possibility that FoxC1 binding might overlap with that of RUNX1 and C/EBPα in FUJIOKA-specific DHSs (Figure 4:3). Our observations that a large proportion of FoxC1 binding events occurred within FUJIOKA-specific DHSs, in the vicinity of C/EBPα or RUNX1, builds upon previous genome co-occupancy patterns determined from DNase I digital footprint bootstrapping analyses in primary FLT3-ITD AMLs, where RUNX, C/EBP, FOX (and FOX:E) footprints formed a tight cluster as compared to other TFs, which did not tend to occupy sites in primary FLT3-ITD AML-specific DHSs[271]. Manual inspection of FUJIOKA-specific DHSs revealed frequent co-occurrence of FoxC1 and C/EBPα binding in ChIP-seq analyses, further corroborating previous DNase-seq digital footprinting analyses.

Collectively, these data indicate that FoxC1 might functionally co-operate with C/EBPα and perhaps also RUNX1 in AML cells to promote the ectopic activation of leukaemia-associated genes. This is consistent with a model proposed by Cauchy *et al.* (2015), where they observed in primary FLT3-ITD AML samples that the normal myeloid RUNX1 and C/EBP binding signature was co-opted in parallel with upregulation of *FOXC1*, leading to the activation of an oncogenic gene signature. Similar roles for other FOX proteins in the retargeting of endogenously expressed TFs have been described elsewhere in oncogenesis, perhaps most convincingly between FoxA1 and the estrogen receptor in breast cancer and FoxA1/FoxO1 with the androgen receptor in prostate

cancer[139,401–403]. However, co-association of binding for more than one TF in independent ChIP-seq experiments is insufficient to confirm a mechanism of co-operative regulation. For this, further experiments are required, as detailed in 4.13.

4.10 The bimodal pattern of FoxC1 binding may be indicative of pioneer activity

The observation that a subset of FoxC1 binding events occurred at regions flanking DHSs – particularly those common to both FUJIOKA and K562 cells – suggests that FoxC1 may instead act as a pioneer factor in this context by binding nucleosomes at target *cis*-regulatory elements to open up nearby chromatin, allowing other factors access to target binding sites. This would be consistent with known activities of other FOX proteins including FoxA1, which was the first documented TF to have true pioneer activity. A distinguishing feature of FOX proteins defined as pioneer factors are their highly context-specific chromatin-binding properties. Binding may occur in a DNA sequence-specific manner, correlating more closely with defined FOX motifs that may exist within DHSs, or they may bind non-specifically to nucleosomes lacking a canonical FOX motif[140,379]. Furthermore, the DNA-binding domain of FOX proteins is the primary structural determinant of these context-specific activities, and this is the most highly-conserved region of FoxC1 as compared to the pioneer factor FoxA1[122,382]. Indeed, the observation that FoxC1 binding within DHSs correlated with the presence of FOX motifs suggests that our ChIP-seq dataset is not simply artefactual.

Finally, it should also be considered that a limitation of the DNase-seq methodology is that it is not well-suited to determining specific nucleosomal positioning patterns, but just detects nucleosome-free regions[404]. Indeed, another high-profile study investigating binding of the pioneer factor FoxA1 with the glucocorticoid receptor found that FoxA1 occupancy was not associated with detectable footprints, further suggesting that DNase I

digital footprinting analyses alone may be an inefficient way to capture FoxC1 binding events that occur in a nucleosomal context. At the time of writing, we have plans for further studies to address current methodological shortcomings, as detailed in 4.13.

4.11 An evaluation of FUJIOKA cells as an alternative cell model for study of FoxC1 function and other targets in primary FLT3-ITD AMLs

Our RNA-seq analyses revealed a modest overlap between differentially regulated genes in FUJIOKA cells as compared to K562 cells and the Cauchy *et al.* (2015) FLT3-ITD-specific gene signature (Figure 4:6 and Table 4:1), including *CST7*, which we demonstrated featured evidence of promoter occupancy by FoxC1 in ChIP-seq analyses (Figure 4:4). Furthermore, comparison of our FoxC1 ChIP-seq data with primary FLT3-ITD AML DNase-seq data revealed that FoxC1 binding was enriched in regions corresponding FLT3-ITD-specific DHSs that were also enriched for FOX:E composite footprints (Figure 4:5).

Collectively, these data suggest that FUJIOKA cells may represent an alternative AML cell line in which to explore mechanisms of gene deregulation which may commonly apply to FLT3-ITD AMLs. In support of this, a very recently published study has identified strong correlations between (a) high *GFI1* expression and the presence of FLT3-ITD, and (b) genes upregulated in the presence of either FLT3-ITD or high GFI expression, even in the absence of FLT3-ITD[405]. High GFI1 expression was associated with high FLT3 expression. Furthermore, siRNA-mediated depletion of *GFI1* in FUJIOKA cells led to the downregulation in surface expression of FLT3 in parallel with several genes upregulated in the Cauchy *et al.* (2015) FLT3-ITD AML gene signature. However, this study did not detect a notable reduction in the growth or viability of FUJIOKA cells following GFI1 suppression, suggesting that this marker was not a survival factor in these cells even though it was associated with the FLT3-ITD gene signature. Moreover, the

phenotypic impact of downregulated FLT3-ITD target genes on the phenotype of FUJIOKA cells is difficult to assess, given the longest time point following *GFI1* depletion was 96 hours. At this time point, depending on factors including protein stability and the efficiency of siRNA knockdown, there may not be a meaningful reduction in protein expression of downstream target genes for a growth or morphological phenotype to become apparent.

Critically, our preliminary shRNA knockdown analyses (Figure 4:8) indicate that FUJIOKA cells may be sensitive to depletion of FoxC1, consistent with FoxC1 knockdown experiments performed in other leukaemia cell lines lacking notable expression levels of FLT3-ITD-specific genes[300,405]. At the time of writing, our analyses so far are encouraging evidence that FUJIOKA cells may be an ideal model in which to explore some outstanding questions which remained from previous studies of FLT3-ITD AMLs conducted in this laboratory[271]. This is significant, because it emerged during the course of this study that two of the most frequently used immortalised FLT3-ITD AML cell lines, MV4-11 and MOLM-14 cells, are poor models for primary FLT3-ITD AMLs, as the chromatin, gene expression and phenotypic signatures of these cell lines are dominated by MLL fusion oncoproteins[406]. The clinical incidence of FLT3-ITD AMLs with *MLL* rearrangements – or indeed any chromosomal translocation - is surprisingly rare[260,407]. The only FLT3-ITD cell line lacking an *MLL* rearrangement, PL-21, instead has a more mature phenotype which is likely due to these cells being derived from an Acute Promyelocytic Leukaemia sample[408]. Indeed, our own chromatin profiling analyses revealed even less overlap between these cells and primary FLT3-ITD AML cells than MV4-11 cells (data not shown).

Finally, we were unsuccessful in obtaining homozygous *FOXC1*$^{-/-}$ clones using a double sgRNA-based CRISPR gene editing approach, which might suggest that this is an

inviable phenotype in FUJIOKA cells. Whilst this is a potentially interesting result in itself, we intend to repeat this experiment to resolve whether apparent inviability is a true biological effect or simply a consequence of the single-cell outgrowth conditions required to generate monoclonal cell lines or low CRISPR efficiency in these cells.

4.12 Conclusions

Through genome-wide chromatin and gene expression profiling, we have identified a specific epigenetic signature present in FUJIOKA cells but not K562 cells. These studies identify a group of almost 15,000 FUJIOKA-upregulated DHSs which were also associated with nearby upregulated genes as compared to K562 cells, a subset of which included commonly activated targets upregulated in primary FLT3-ITD AML patients. Furthermore, these DHSs were defined by an enrichment of motifs and binding for the TFs RUNX1, C/EBPα and FoxC1, in concordance with previous genome-wide occupancy analyses of primary FLT3-ITD AML cells.

Collectively, these findings suggest that FUJIOKA cells indeed share common features of gene deregulation with primary FLT3-ITD AMLs. Crucially, given the absence of stable FLT3-ITD cell lines relevant to the overwhelming majority of primary human FLT3-ITD AMLs, FUJIOKA cells may represent a suitable alternative model in which to advance our understanding of the impact some FLT3-ITD-activated genes, including *FOXC1*, have on the development of leukaemia in primary AML patients.

Furthermore, analyses of FoxC1 binding patterns in FUJIOKA cells revealed an unexpected bimodal binding signature that may be indicative of pioneer activity, as well as potential co-operativity with C/EBPα in the activation of *cis*-regulatory elements associated with oncogene activation.

Our analyses represent a starting point for further mechanistic interrogation of FoxC1 function in the context of AML. An aim of future work in this laboratory is to eventually extend these studies to a patient-derived xenograft mouse model of primary FLT3-ITD AML, which was only recently developed in collaboration with the Heidenreich laboratory.

4.13 Current limitations and future prospects

Our studies of FUJIOKA cells are so far almost entirely descriptive, and need to be followed up with functional experiments to validate our hypotheses. At the time of writing, we are repeating *FOXC1* CRISPR deletion experiments in an attempt to delete the second *FOXC1* allele. If the *FOXC1*$^{-/-}$ cells are viable, these loss of function experiments should help to reveal FoxC1-dependent changes in gene regulation by comparison with our existing datasets. These analyses will be complemented further by shRNA-mediated knockdown experiments, firstly for *FOXC1* and subsequently for putative target genes. Significantly, we plan to use these knockdown experiments as a platform from which to extend the FoxC1 studies into primary human AML cells expanded using a patient-derived xenograft mouse model. Using this model, we will finally be able to confidently link FoxC1-dependent changes in chromatin features and gene expression to phenotypes in a more directly relevant model of human AMLs.

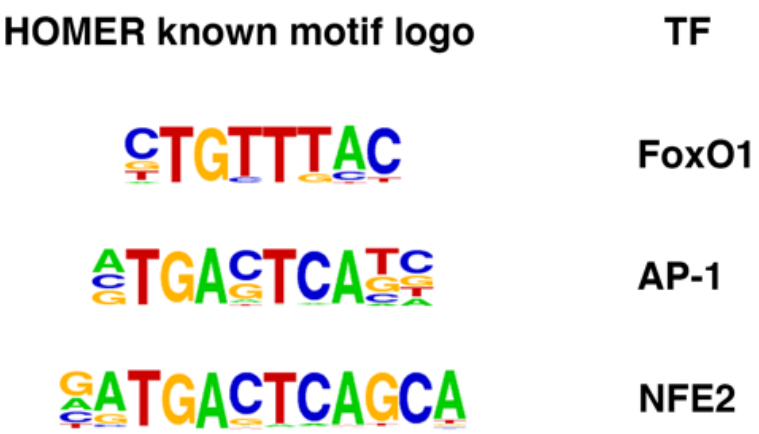
Our ChIP-seq analyses gave preliminary indications of a functional co-operativity between FoxC1 and other haematopoietic TFs in the context of AML, but require further experiments to increase confidence that this is a true biological effect. To explore this, we could perform FoxC1 co-immunoprecipitation experiments followed by mass spectrometry to identify interacting partners of FoxC1 in an unbiased manner. Next, the contribution of candidate interactors identified could be assessed using RNAi-based knockdown approaches paired with further ChIP-seq studies and/or chromosome

interaction assays, to evaluate changes in *cis*-element occupancy and potential promoter-distal element looping.

Next, questions remain as to the mechanism by which FoxC1 binds to target sites in FUJIOKA cells. The observed bimodal pattern of binding, either close to DHS summits enriched for FOX motifs or in nucleosomal regions that flanked DHSs, was indicative of possible pioneer activity. However, several experiments are required to test this possibility further. Firstly, these ChIP-seq analyses should be repeated, ideally with an independent antibody, to eliminate the possibility that this is an artefact of our analyses. However, given that we independently observed a similar trend in a ChIP-seq dataset prepared by another lab in mice (Chapter 3), we consider this possibility to be unlikely. Furthermore, as a defining feature of some FOX proteins are their ability to bind to chromatin independently of DNA sequence[140,379], it would be crucial to assess binding of FoxC1 directly to nucleosomes. To this end, we are now preparing His-tagged FoxC1 expression constructs to generate recombinant protein in *E. coli* for purification and use in nucleosome binding assays.

Finally, a limitation of our current analyses stems from the DNase-seq methodology, which efficiently detects nucleosome-free regions but does not precisely map alterations in nucleosome positioning. Using this method in isolation, we are unable to identify regions where FoxC1 might bind nucleosomal sites. To resolve this outstanding question, we could perform titrated MNase-seq, which has successfully been used to discriminate between FoxA binding at both accessible and nucleosomal sites[379]. Integrating our ChIP-seq data with the MNase-seq data would allow us to understand how FoxC1 binding relates to underlying nucleosome structure, and would complement *in vitro* nucleosome binding data. Furthermore, MNase-seq analyses could be performed following FoxC1 knockdown and compared to untreated FUJIOKA cells, to investigate

FoxC1-dependent shifts in nucleosome positioning. Collectively, these *in vitro* and genome-wide analyses would provide robust indicators of whether FoxC1 is indeed a pioneer factor or not.



Appendix 4:1 **Known TF binding motif logos from the HOMER database used in these analyses.**

As described in this chapter, some motif analyses involved known motif searches, specifically the FoxO1 motif, a validated canonical FOX binding site, and the NFE2 motif.

The known AP-1 motif in the HOMER database is shown for comparison to the NFE2 motif.

Chapter 5: Ectopic expression of *FOXC1 is* associated with the activation of FOX:E composite *cis*-regulatory elements in primary *FLT3*-ITD AML patients.

Immediately prior to beginning these studies, ongoing work in this laboratory determined that primary FLT3-ITD AML samples were distinguished from other primary AML samples by a chromatin signature featuring the specific activation of FOX:E (FOX:E) composite elements. These data were consistent with the binding of FOX family proteins co-ordinately with basic helix-loop-helix (bHLH) family E-box-binding proteins, in parallel with the FLT3-ITD-specific upregulation of *FOXC1* mRNA expression[271]. In this chapter, I will introduce FLT3-ITD-specific FOX:E *cis*-regulatory elements, then explore their function in order to understand their activity in the context of FLT3-ITD-dependent gene activation.

5.1 FLT3-ITD AML samples feature the specific activation of *FOXC1* and DHSs enriched for FOX or FOX:E composite binding motifs

It was previously introduced in 1.5.6 that *FOXC1* is frequently upregulated in AML patient samples with FLT3-ITD (Figure 5:1A). Furthermore, activation of the *FOXC1* locus was associated with a recurrent FLT3-ITD-specific DHS 56 kb upstream of the promoter that was enriched for RUNX motifs and demonstrated by ChIP-seq analyses to be occupied by RUNX1 (Figure 5:1B and C). As these patients also featured elevated *RUNX1* expression as compared to FLT3 wild-type (FLT3-WT) patients, it was postulated that RUNX1 may be involved in the ectopic activation of *FOXC1* in this context[271].

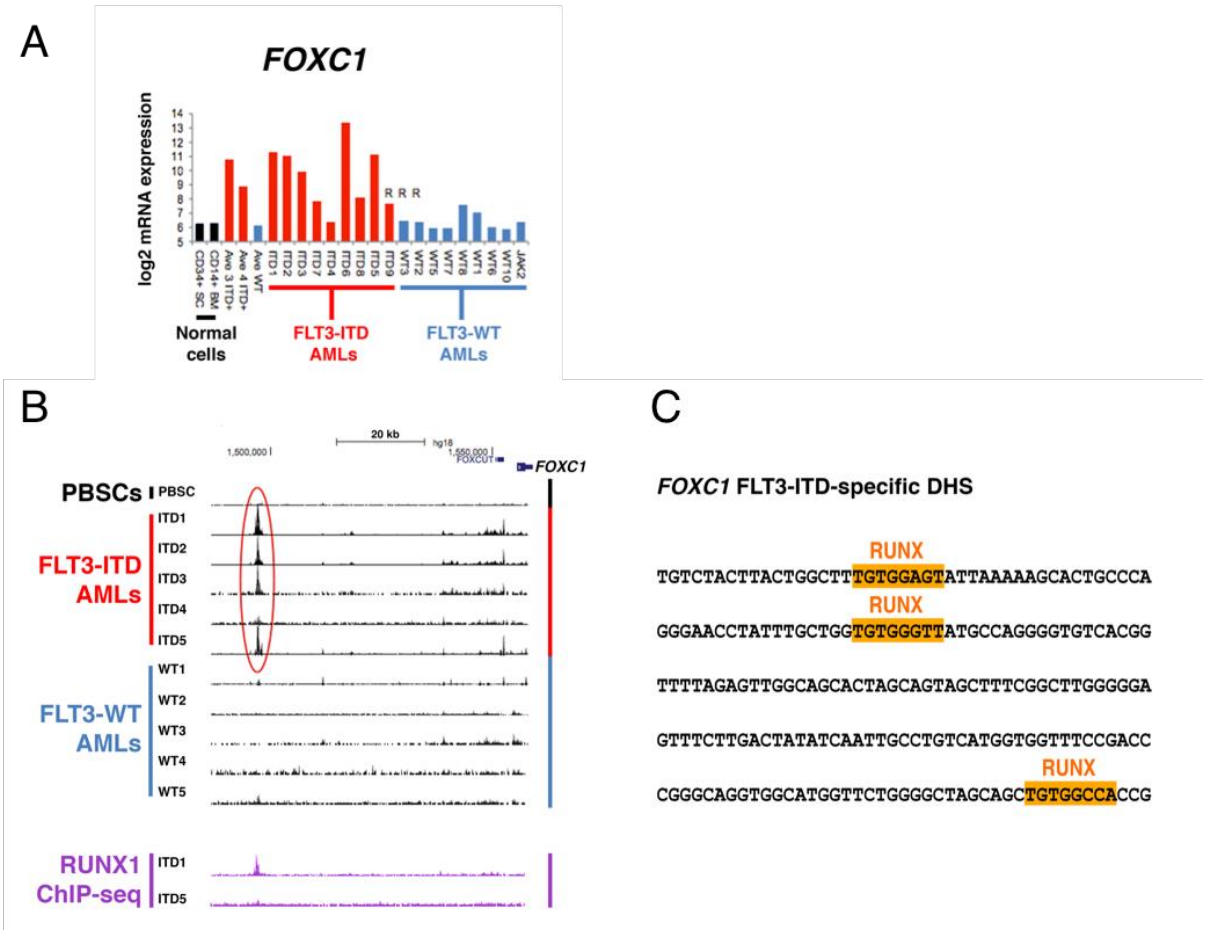


Figure 5:1 **FLT3-ITD-dependent activation of *FOXC1* is associated with a differential DHS enriched for RUNX motifs and RUNX1 occupancy**

A: Analyses of normalised microarray mRNA data of AML patient samples and normal cells from Cauchy *et al.* (2015) for *FOXC1*. Primary FLT3-ITD AML samples are coloured in red, whereas wild-type FLT3 samples are coloured in blue. Normal CD34+ PBSCs (CD34+ SC) and CD14+ BM monocytes (CD14+ BM) are coloured in black. AML samples carrying *RUNX1* mutations are indicated with an **R** above their mRNA expression values.

B: UCSC genome browser view of DNase-seq data (upper tracks) for the human *FOXC1* locus. The -56 kb FLT3-ITD-specific DHS is circled in red and also features RUNX1 binding in the anchor sample for this study, ITD1, as shown by RUNX1 ChIP-seq data (at bottom).

C: Genomic sequence detail of the -56 kb *FOXC1* upstream DHS circled in **B**, with RUNX binding motifs highlighted.

Further inspection of microarray analyses published in this study revealed that *FOXC1* was the only substantially upregulated FOX gene in FLT3-ITD AML patients as compared to all FOX genes expressed in primary AMLs with/without FLT3-ITD, and normal CD34+ PBSCs (Figure 5:2A). However, this pattern was not consistent for AML

cell lines, which were found by similar analyses to feature only modest levels of *FOXC1* expression (Figure 5:2B).

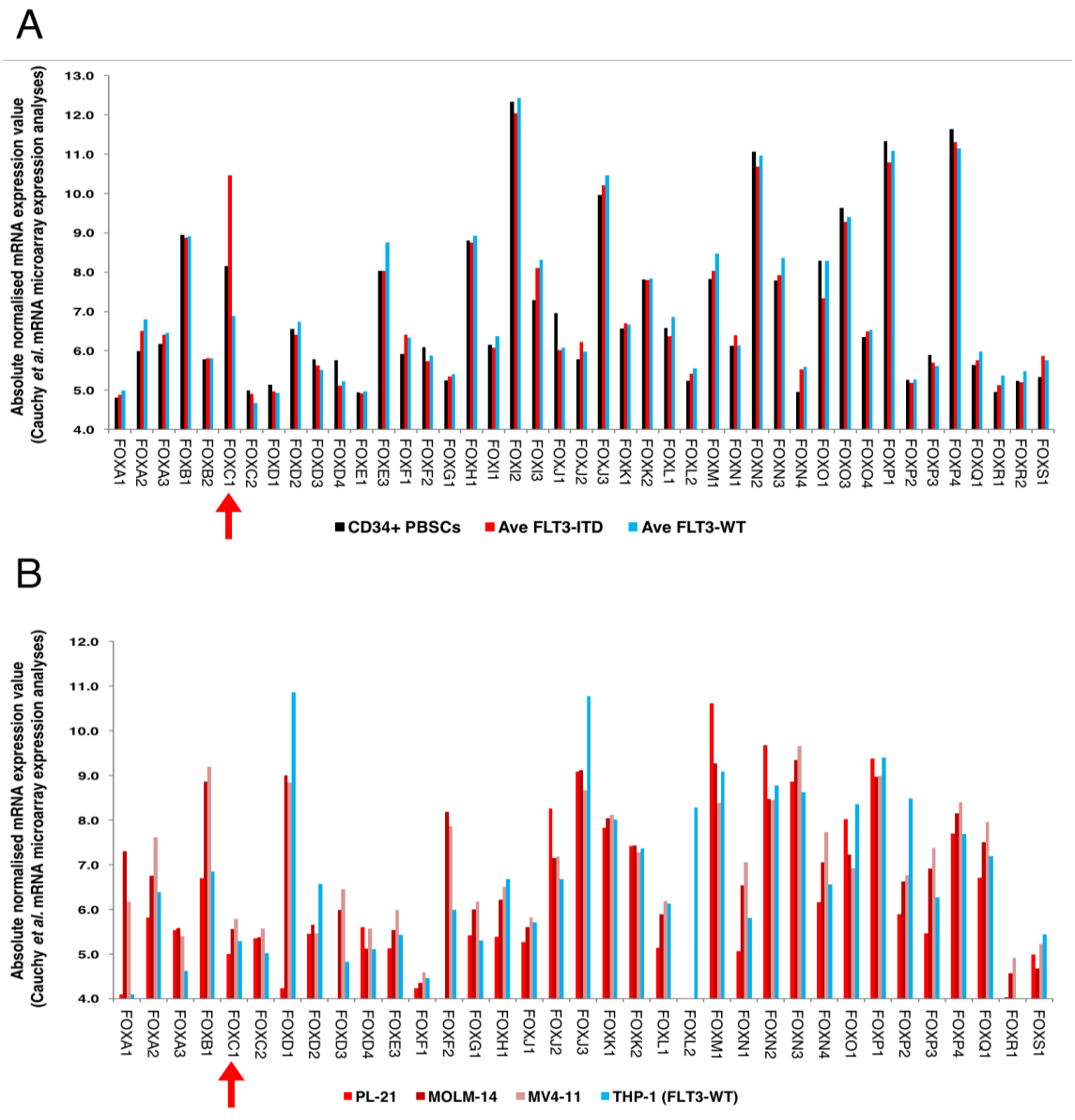


Figure 5:2 ***FOXC1* upregulation in FLT3-ITD AMLs is specific compared to other genes in the FOX family.**

A: Analyses of primary AML and CD34+ PBSC mRNA expression data from Cauchy *et al.* (2015) reveal FLT3-ITD-specific upregulation of gene expression of *FOXC1*.

B: Microarray analyses reveal the specific complement of FOX gene mRNAs expressed in three FLT3-ITD AML cell lines (PL-21, MOLM-14 and MV4-11), together with the FLT3-WT cell line, THP-1.

In both panels, *FOXC1* mRNA expression is indicated underneath with a red arrow.

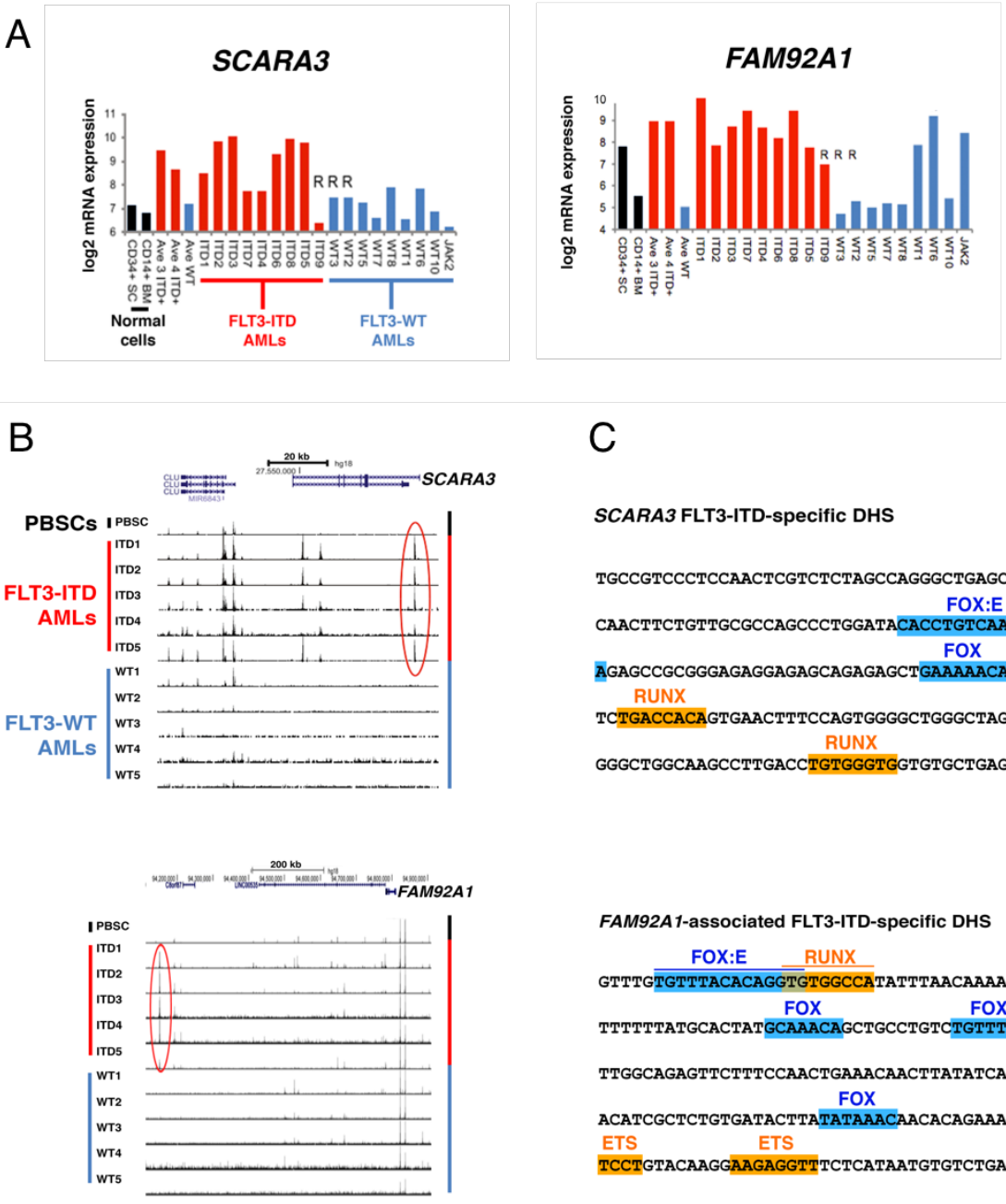


Figure 5:3 ***SCARA3* and *FAM92A1* are two examples of FLT3-ITD-specific DHSs distinguished by FOX:E composite sites**

A: Analyses of normalised microarray mRNA expression values for *SCARA3* and *FAM92A1* from primary AML samples, presented as in Figure 5:1.

B: UCSC genome browser views of FLT3-ITD-specific DHSs associated with the *SCARA3* and *FAM92A1* loci (circled in red).

C: Genomic sequence detail for *SCARA3* and *FAM92A1* DHSs, with FOX:E composite site highlighted together with other TF binding motifs.

In addition to FLT3-ITD-specific upregulation of *FOXC1*, this study also identified several differentially-activated DHSs enriched for FOX and FOX:E composite binding motifs in FLT3-ITD AML samples. Furthermore, many of these DHSs were associated with the upregulation of nearby FLT3-ITD-specific target genes, including *SCARA3* and *FAM92A1* (Figure 5:3). Finally, the observation that many of these FOX:E sites were predicted to be occupied *in vivo* by high-resolution DNase-seq digital footprinting analyses (Figure 5:4A) strongly suggested that DHSs containing FOX:E elements might play a critical role in the FLT3-ITD-dependent activation of target genes promoting leukaemogenesis. Thus, the aims of these studies were to delineate mechanisms of FOX:E element activation by TFs and whether these sites functioned as a novel class of enhancer elements to drive expression of FLT3-ITD target genes.

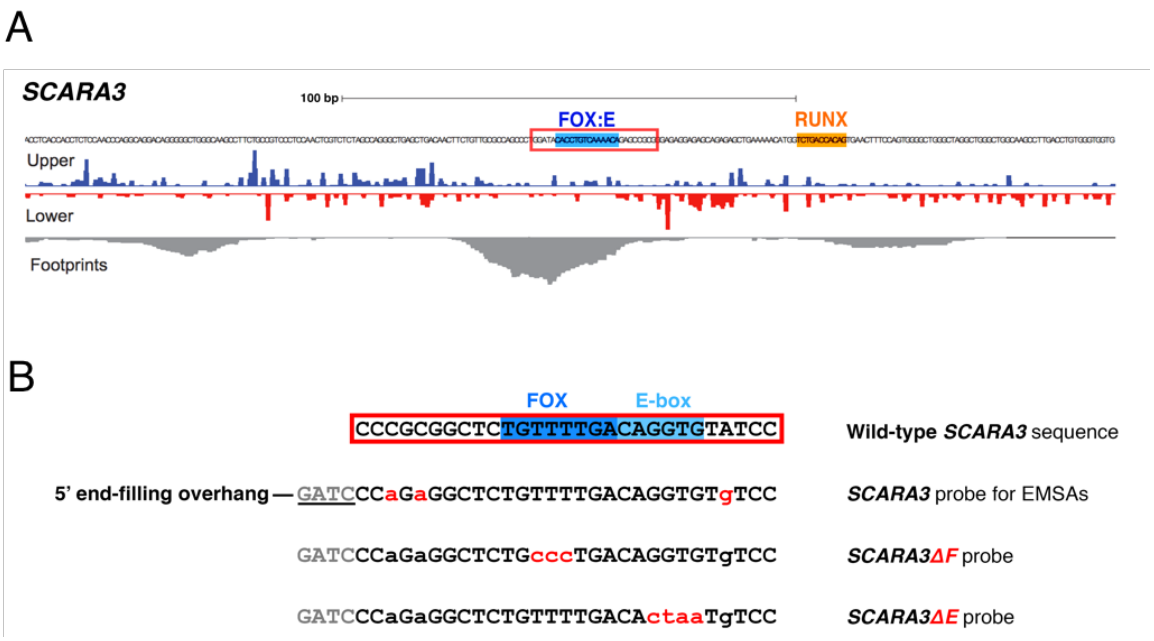


Figure 5:4 **The *SCARA3* FLT3-ITD AML-specific DHS features FOX:E footprints *in vivo*.**

A: (Top) DNA sequence of the human *SCARA3* FLT3-ITD AML-specific DHS. Underneath are the strand-specific DNase I cutting patterns in a representative FLT3-ITD AML patient sample, as revealed by very high depth sequencing. Both the FOX:E composite element and a RUNX site are preferentially protected from nuclease digestion. At the bottom of this panel are the predicted occupancy (footprint) probabilities as determined by from the DNase cutting patterns, with higher confidence footprints

presented as taller peaks (e.g. the FOX:E footprinted site) as compared to lower confidence footprints.

B: View of the wild-type *SCARA3* FOX:E composite site and flanking sequences used for the design of probes in EMSA assays. The 'intact' *SCARA3* sequence used in EMSAs features three point mutations to ablate a low-affinity GC-box and GATA site to improve specificity of binding reactions (changed residues in red), as well as an additional 5' overhang sequence used for end-filling probes with radiolabelled cytosine (^{32}P-dCTP). Mutant sequences lacking FOX or E-box sites were also tested.

5.2 Observation of E-box binding to FOX:E composite elements *in vitro*

To examine whether FOX:E elements were occupied by TFs recognising the FOX and E-box sequences *in vitro*, we performed EMSA assays using probes containing FOX:E composite elements specifically activated in primary FLT3-ITD AML samples (Figure 5:4B). For these analyses, the prototype FOX:E sites associated with *SCARA3* and *FAM92A1* loci were tested the most extensively.

Firstly, evidence of potential E-box-binding was observed in the form of three discrete bands with both the *SCARA3* and *FAM92A1* probes when using nuclear extracts prepared from MV4-11 cells, which express FLT3-ITD in addition to the MLL-AF4 fusion oncoprotein (Figure 5:5). The specificity of this binding to both probes was confirmed using competition EMSA assays, where a 100-fold molar excess of various unlabelled competitor E-box or FOX:E sequences were all sufficient to disrupt E-box binding. However, no evidence of specific binding to the FOX site in these probes was observed with MV4-11 nuclear extracts. A similar pattern of binding was observed in EMSAs performed with MOLM-14 cells, which also feature FLT3-ITD in addition to the MLL-AF9 fusion oncoprotein (Figure 5:8).

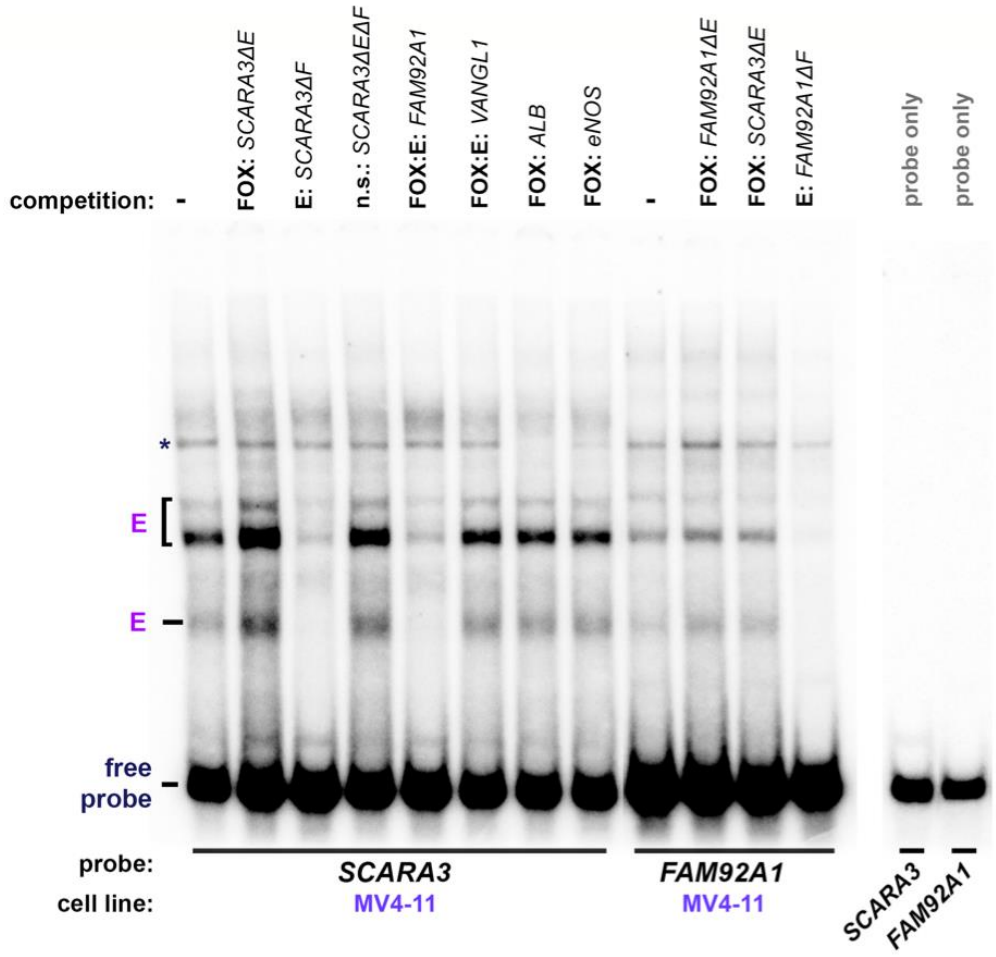


Figure 5:5 **Formation of specific E-box complexes at FOX:E composite elements using MV4-11 nuclear extracts.**

EMSAs of *SCARA3* and *FAM92A1* FOX:E composite elements identified in Cauchy *et al.* (2015) as activated in FLT3-ITD AML patients. Binding reactions contained nuclear extract from MV4-11 cells, which carry FLT3-ITD in addition to an *MLL* translocation.

Specificity of E protein complexes (**E**) were tested using a 100-fold molar excess of homologous or heterologous unlabelled DNA competitor sequences (indicated at top). (* = non-specific band)

5.3 Evidence of specific FOX and E-box motif occupancy *in vitro*

Because no evidence of specific FOX motif binding was observed in EMSAs with either MV4-11 or MOLM-14 cells, we next performed EMSAs with PL-21 cells. Like MV4-11 and MOLM-14, PL-21 cells carry FLT3-ITD and express modest levels of *FOXC1* mRNA (Figure 5:2), but importantly do not feature an *MLL* translocation.

In contrast to observations made with other FLT3-ITD cell lines, EMSAs performed using nuclear extracts from PL-21 cells did exhibit signs of specific FOX motif binding in addition to E-box binding (Figure 5:6). FOX motif binding was distinguished by the presence of an additional band with a particularly slow electrophoretic mobility, appearing in the upper region of gel images. In contrast to the E-box binding complexes, which varied in mobility when bound either to the *SCARA3* and *FAM92A1* probes, FOX motif binding caused a band shift to the same position on the gel, regardless of which probe was used. Furthermore, competition assays using a range of unlabelled FOX and FOX:E DNA competitors suggested putative FOX binding to both probes was specific.

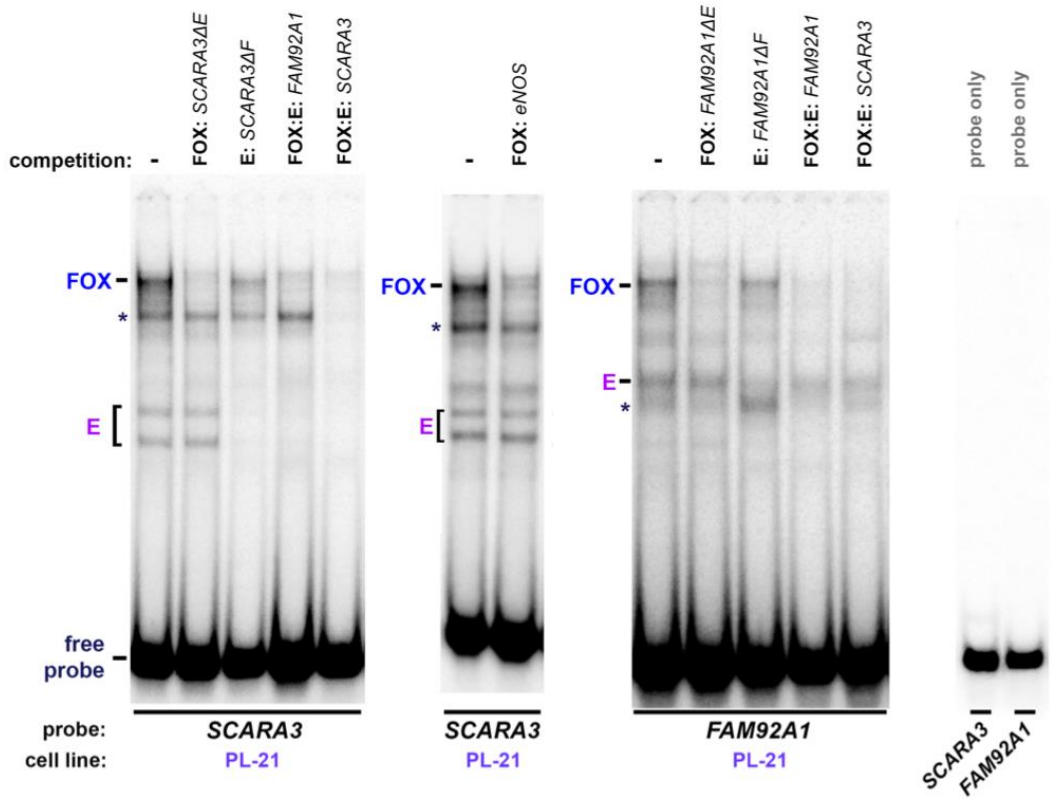


Figure 5:6 **Evidence of putative FOX complexes at *SCARA3* and *FAM92A1* FOX:E composite elements using PL-21 nuclear extracts.**

Evidence of E protein (**E**), and FOX protein (**FOX**) protein complexes are indicated to the left of gel images. Specificity of complexes was tested using a 100-fold excess of unlabelled DNA competitor as in Figure 5:1. (* = non-specific band)

To further verify specificity of FOX protein binding, we tested the *SCARA3* probe in EMSAs with nuclear extracts prepared from THP-1 cells. Although THP-1 cells do not carry FLT3-ITD, they are known to express moderate levels of FoxC1 protein, in contrast to most other available AML cell lines[300]. Indeed, we found evidence of specific binding of both FOX-like proteins and E-box–binding bHLH-like protein binding between THP-1 nuclear extracts and the *SCARA3* probe (Figure 5:7). The binding of FOX-like proteins to the *SCARA3* probe was efficiently disrupted using homologous FOX competitor sequences and a range of other previously published sequences that are known to specifically bind the FOX family proteins FoxO1 (*eNOS*), FoxA1 (*ALB*) and FoxA2 (*CLPS*)[141,409,410].

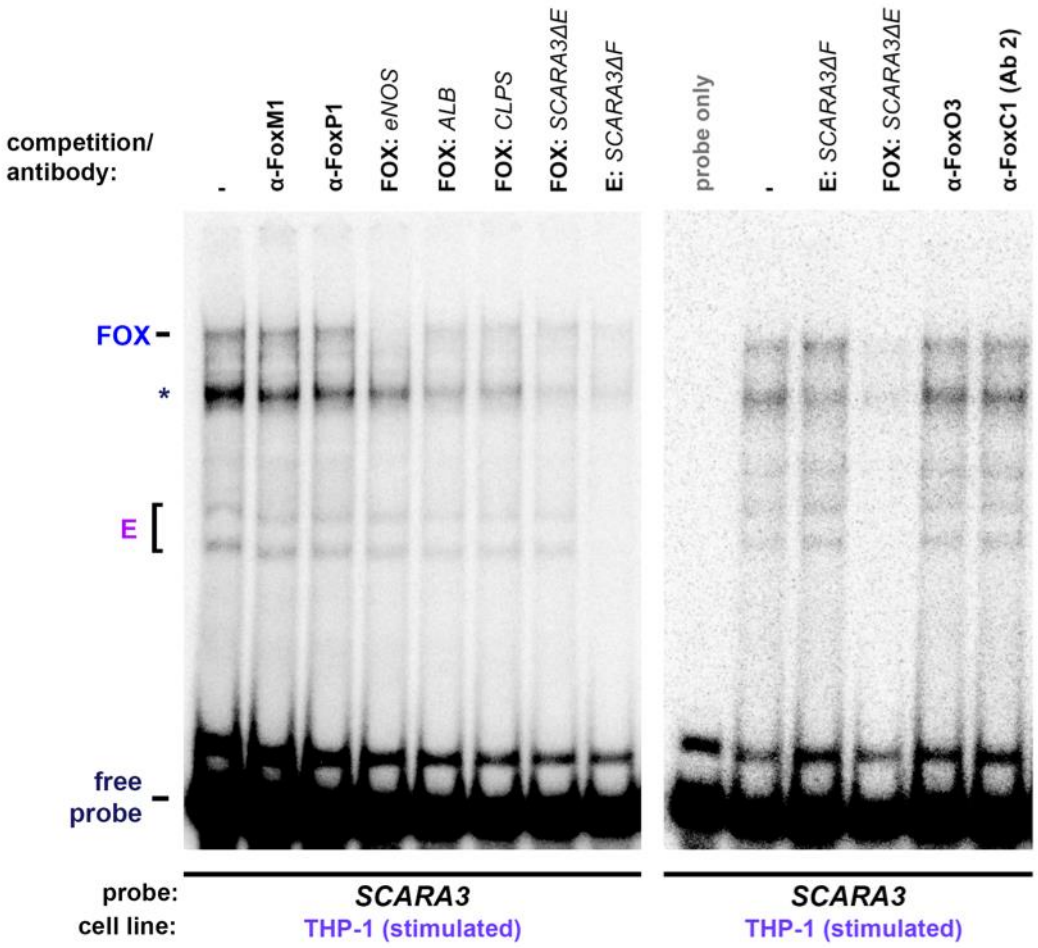


Figure 5:7 **Further competition experiments support FOX protein binding to the *SCARA3* probe, but supershift assays fail to identify the specific FOX family member involved.**

EMSAs performed on stimulated THP-1 cell nuclear extracts with the *SCARA3* probe and further heterologous FOX competitor sequences or α-FOX protein antibodies.

In contrast to *FAM92A1*-based FOX competitors, several other heterologous competitors efficiently disrupted FOX protein binding with the *SCARA3* probe, including those from three published studies of the FOX proteins FoxO1 (*eNOS*), FoxA1 (*ALB*) and FoxA2 (*CLPS*).

However, further α-FOX protein antibodies tested did not alter the formation or migration of putative FOX:*SCARA3* complexes. (* = non-specific band)

Finally, we sought to further validate specificity of E-box and FOX-like protein binding in EMSAs by using antibodies raised against various bHLH and FOX proteins, including FoxC1, but were unsuccessful in disrupting or altering the mobility of FOX-like complexes using this approach (Figures 5:7 and 5:8). In parallel with one of these

experiments (Figure 5:8), we tested the competitive activity of a third putative FOX:E composite element present at the *VANGL1* locus. These assays revealed that this sequence was a poor competitor of both E-box or FOX motif binding at the *SCARA3* probe.

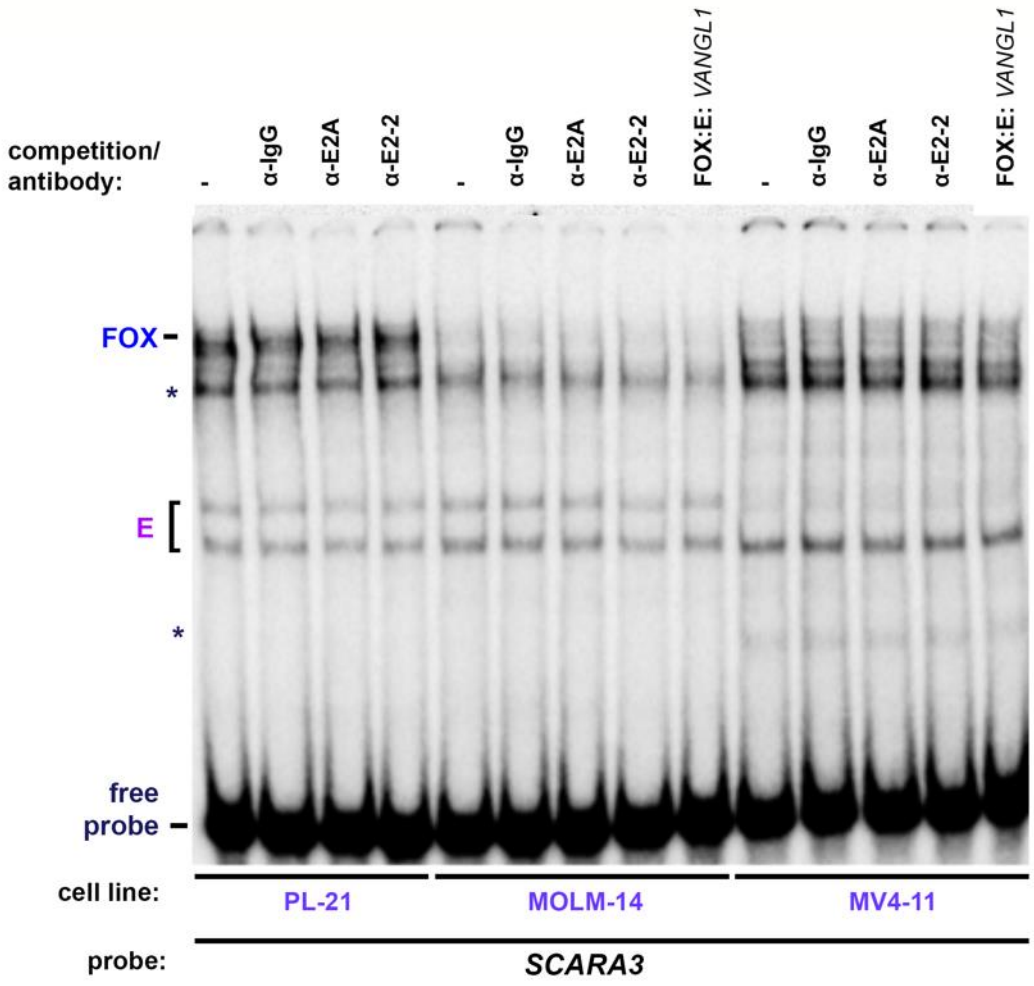


Figure 5:8 **Supershift EMSA assays fail to identify specific E-box binding proteins in various nuclear extracts.**

EMSAs performed on nuclear extracts prepared from three FLT3-ITD AML cell lines in combination with antibody raised against two known E-box binding proteins, E2A and E2-2, or a third putative FOX:E composite site at the *VANGL1* locus. (* = non-specific band)

5.4 FOX and bHLH protein binding to the *SCARA3* FOX:E element is not enhanced following cell stimulation

Because many TFs including FOX proteins exhibit signalling-dependent changes in DNA-binding activity, we sought to test whether observed TF binding in EMSAs was affected by stimulation of signalling pathways which might target FOX and bHLH family proteins for post-translational modification. For these experiments we used THP-1 cells, which already exhibited specific FOX and E-box motif binding in previous EMSAs, but are also known to feature inducible cell signalling following treatment with PMA and Calcium Ionophore[411,412].

At the *SCARA3* probe, we detected no appreciable difference in the extent of FOX or E-box binding using unstimulated or stimulated THP-1 nuclear extracts (Figure 5:9). However, previous evidence of FOX and E-box disruption using competitor sequences was reproducible regardless of whether the nuclear extract used was prepared from stimulated or unstimulated cells.

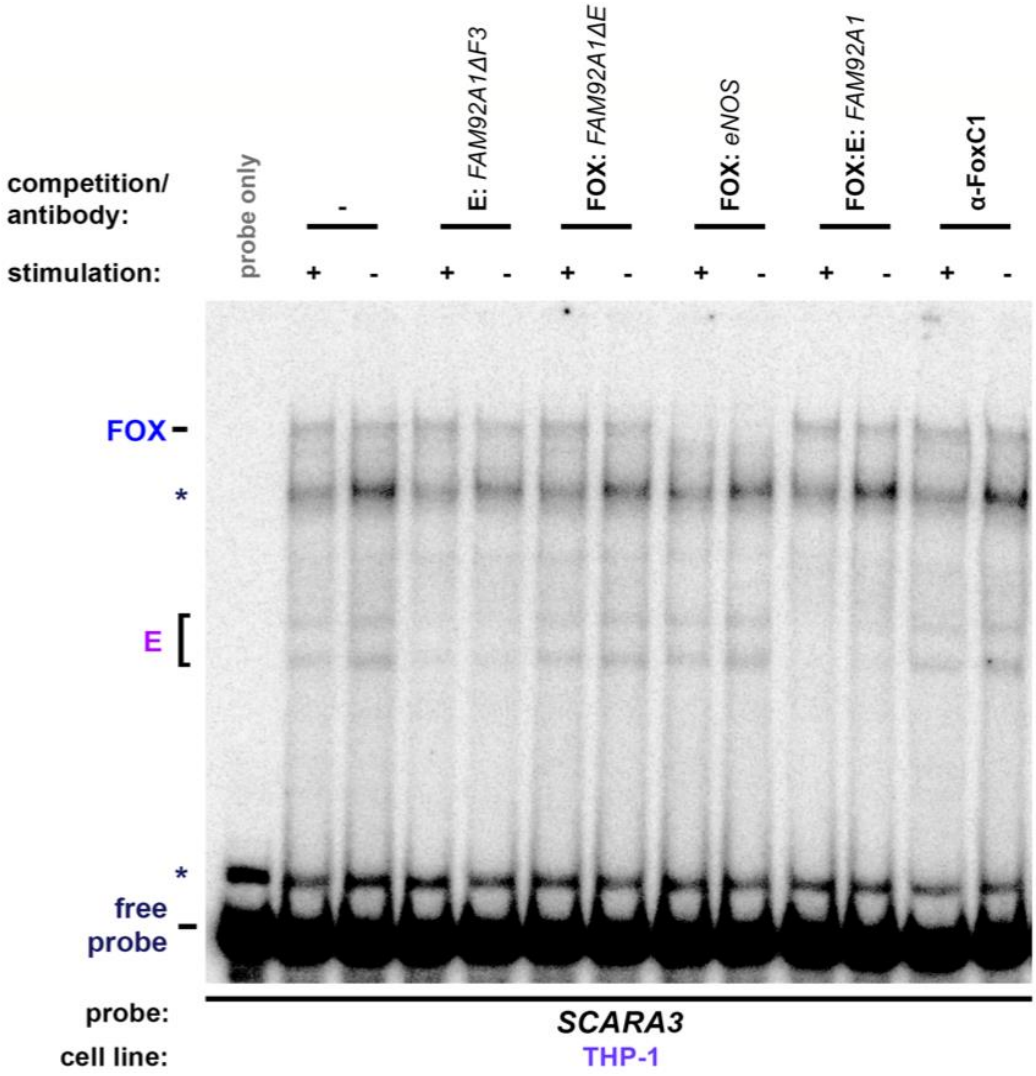


Figure 5:9 **FOX and E protein binding is not enhanced following stimulation of mitogenic and inflammatory cell signalling in THP-1 cells.**

EMSAs performed on nuclear extracts from THP-1 leukaemia cells, with/without stimulation of mitogenic and inflammatory signalling using PMA and Calcium Ionophore.

Specific **FOX** and **E** protein complexes formed with the *SCARA3* probe, which was supported by competition assays as in figures 5:1 and 5:2. Critically, *FAM92A1*-based competitor sequences did not efficiently outcompete putative FOX binding to the *SCARA3* probe.

However, there were no observable differences in the intensity or composition of specific complexes forming with/without stimulation, except at the upper non-specific band (*) migrating between FOX and E protein bands. (* = non-specific band)

5.5 FLT3-ITD-specific DHSs containing FOX or FOX:E composite sites do not function as classical enhancer elements as tested in luciferase reporter assays

Given that we observed evidence of specific FOX and E-box occupancy at FOX:E sequences *in vitro*, we sought to test whether DHSs containing FOX or FOX:E composite sites acted as enhancer elements when tested in a Firefly luciferase reporter construct containing a minimal promoter fragment[326]. For these studies we used FUJIOKA cells, which feature robust FoxC1 expression and are amenable to plasmid co-transfection as required by this assay. Because primary FLT3-ITD AML samples are distinguished by chronic MAP kinase signalling activation, we also tested the enhancer output of these reporter constructs in Fujioka cells stimulated with PMA and Calcium ionophore for 8 hours.

Of the 7 constructs tested containing DHS fragments with FOX or FOX:E sites, none enhanced the ability of a minimal promoter to drive expression of a Firefly luciferase reporter gene as compared to a promoter-only control or positive enhancer control (Figure 5:10). The same trend was observed for an additional pair of reporter constructs, featuring either 3 copies of the *FAM92A1* FOX:E composite site or a fragment of the *FOXC1* -56 kb upstream regulatory element specifically activated in primary FLT3-ITD AML samples (FOXC1). Furthermore, none of the constructs tested featured significantly increased luciferase reporter activity in cells stimulated with PMA and Calcium ionophore. Notably, the *IL2RA*(2) DHS fragment was substantially induced following FUJIOKA cell stimulation but not to a point that was statistically significant ($p > 0.05$) when compared to the uninduced conditions via a paired t-test. This was in stark contrast to the well studied SV40 enhancer, which supported a high level of enhancer activity[413,414].

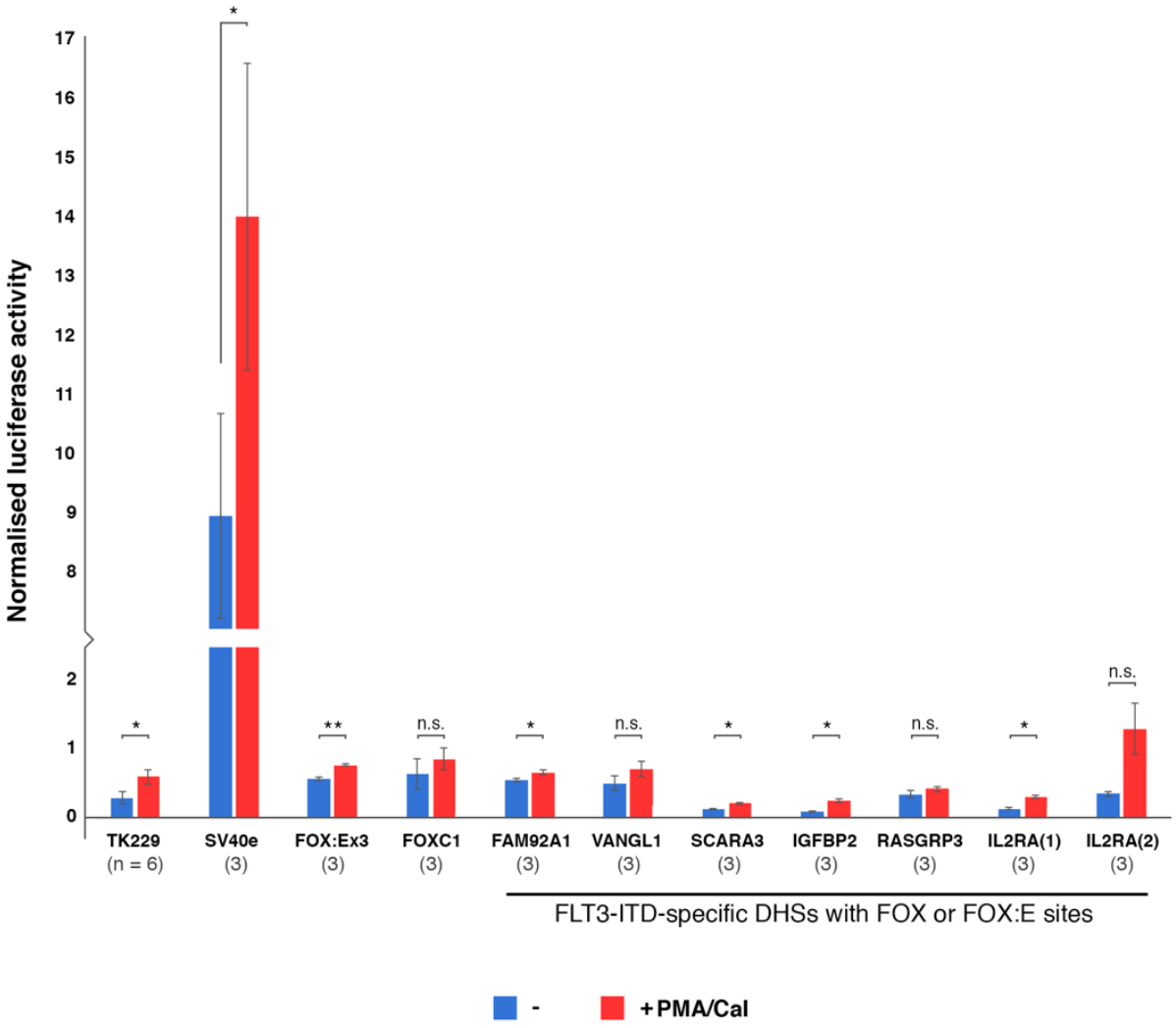


Figure 5:10 **DHSs containing FOX:E composite elements do not operate as classical enhancers in FUJIOKA cells as tested in luciferase reporter assays.**

Luciferase reporter assays for various Firefly luciferase reporter constructs containing either 3 copies of the *FAM92A1* FOX:E site (FOX:Ex3), fragments of DHSs containing FOX:E motifs, or a FLT3-ITD AML-specific DHS associated with activation of the *FOXC1* locus itself (FOXC1) as tested in FUJIOKA cells under non-induced (-) or stimulated (+ PMA/CaI) conditions.

Assays are presented as the average of several biological replicates normalised against in internal Renilla luciferase reporter control construct, and compared to a promoter-only (TK229) and positive enhancer (SV40e) control contructs.

* = p <0.05; ** = p <0.00; n.s. = not significant (p ≥0.05)

Chapter 5: Discussion

5.6 Despite evidence of specific FOX and E-box motif binding *in vitro*, the specific factors occupying FOX:E elements remains unclear

In this chapter, we explored the potential roles of FOX:E DNA elements in the regulation of AML cells. We were able to observe the presence of potential FOX-like and bHLH protein-like complexes forming specifically with FOX and E-box DNA motifs respectively in some of the nuclear extracts tested, particularly those prepared from PL-21 cells and THP-1 cells. Using a range of homologous and heterologous unlabelled competitor sequences, including those with specific FOX or E-box sites ablated by mutations, we were able to determine with reasonable certainty that these sites were specifically occupied *in vitro*.

Our observation of two E-box protein-like complexes is consistent a known mechanism of DNA binding by bHLH proteins, which involves each monomer of a homo- or hetero-dimeric HLH protein complex binding to each half site of the palindromic CANNTG E-box sequence[415]. Thus, it is possible that one complex represents a homodimer of a protein such as E2A, while the other complex may involve a heterodimer such as E2A/TAL1. We were unable, however, to confirm the identities of any of the HLH protein-like complexes using any of the HLH antibodies that we employed. FOX DNA motif binding complexes appeared as a single specific band, which agrees with other EMSA data with confirmed FOX protein binding to free DNA[141,147]. We were, however, also unable to confirm the identities of the FOX protein-like protein complexes. There are several possibilities that might explain this. Firstly, we illustrated in Figure 5:2B that of the ~43 human FOX genes classified to date, over 30 of these were abundantly expressed at the mRNA level in the AML cell lines used to prepare nuclear extracts for testing in EMSAs. We therefore

tested available antibodies raised against four candidate FOX genes, FoxC1, FoxM1, FoxO3 and FoxP1, identified as expressed in analyses of microarray data or by Western blot (Figure 5:2 and Appendix 5:1), together with antibodies against the E-box binding proteins, E2-2 and E2A. These included two antibodies raised against FoxC1, the candidate FOX protein binding to FOX:E motifs *in vivo* in FLT3-ITD AMLs.

We considered the possibility that these antibodies were inefficient at binding target epitopes, and tried various alternative protocols to our standard EMSA protocol to overcome this. Firstly, we employed an extended binding reaction time of 1 h at 4°C prior to gel electrophoresis, to allow greater opportunity for antibody to conjugate with target proteins. Next, we tried varying quantities of antibody and non-radioactive polydI-dC to shift binding kinetics in favour of antibody-protein binding. Finally, we changed the order in which components of the binding reaction were added, incubating nuclear extract with the antibody prior to addition of the EMSA probe. Ultimately, all of these modifications still failed to cause any detectable disruption or electrophoretic mobility shift to FOX motif binding (data not shown), leading us to surmise that despite reasonable evidence correlating *FOXC1* expression with the activation of FOX:E elements *in vivo*, the reagents and tools available at the time of this analysis were not appropriate to confirm this *in vitro*. Having invested over 18 months in these optimisation experiments, and considering that there were dozens of both FOX and bHLH proteins expressed in AML cells, we decided that it would be impractical and financially inviable to test every available antibody. Consequently, the overall focus of this study shifted to characterisation of other molecular changes related to FoxC1 upregulation, as discussed in the previous two results and discussion chapters.

Ultimately, these redirected efforts proved to be a more productive use of limited research time, despite these outstanding questions. We obtained more robust evidence

of FOX protein activity using these alternative approaches, particularly high-resolution DNase-seq analyses in the Hoxa9+FoxC1 mouse AML model, and ChIP-seq analyses of FoxC1 binding in FUJIOKA cells (Chapters 3 and 4). We are now in a position to revisit the above *in vitro* DNA binding studies using alternative approaches, detailed in section 5.9.

5.7 FOX:E composite elements may act as priming DHS elements in FLT3-ITD AMLs

In an effort to obtain direct evidence of the function of DHSs containing FOX:E motifs in the activation of transcription, we tested fragments of FOX:E DHSs in luciferase reporter gene assays. Although we succeeded in establishing a reliable reporter assay in FUJIOKA cells, none of the FOX:E motif-containing DHS fragments we tested enhanced transcription as compared to the SV40 enhancer control, which caused a substantial, reproducible increase in luciferase signal.

One possibility that may account for these observations is that FOX proteins targeting FOX:E elements may not efficiently recognise target sites as they exist on supercoiled plasmids in reporter gene assays. Although it has previously been demonstrated that transfected plasmids acquire some degree of chromatinisation in mammalian cells[416], there is still dispute over whether the chromatin topology of these sequences reliably reflects the native sequence as it would exist in native chromatin[417]. This may be relevant in relation to occupancy of FOX:E sites, since in other studies binding of FOX proteins to target loci appears to be at least partly dependent upon the chromatin context. For example, FoxA1 has been demonstrated to “scan” chromatin prior to recognition of target sites and stable binding[140,150]. Thus, it is possible that FOX proteins may not recognise binding sites as they exist on plasmids with an atypical chromatin structure. To test this hypothesis, an alternative enhancer reporter assay interrogating sequences of interest as

they exist in native chromatin is required. One appropriate strategy to test FOX:E sites is using site-specific integration FACS followed by sequencing (SIF-seq), which allows medium throughput screening of *cis*-regulatory elements in an unbiased manner[418]. However, a more compelling possibility that may explain our observations is that DHSs containing FOX:E composite elements may instead act as priming DHSs (pDHSs) rather than as classical enhancers. Previous studies published by our group defined pDHSs in T cells as regions of accessible chromatin that stably maintain an active chromatin structure at inducible genes without directly activating transcription themselves[184]. We determined in these studies that pDHSs lack classical enhancer activity as tested in reporter gene assays, similar to the results obtained here with DHSs containing FOX:E sites.

An additional defining feature of pDHSs defined in T cells was their close proximity to highly inducible DHSs strongly enriched for motifs recognised by the inducible factors NFAT and AP-1, whereas pDHS are most highly enriched for ETS and RUNX motifs. This is of relevance to our observations, given that the FLT3-ITD-specific DHSs identified in Cauchy *et al.* (2015) were also most strongly enriched for ETS and RUNX binding sites, including those DHSs containing FOX:E motifs. Furthermore, our observations that binding to FOX and E-box motifs in EMSAs was not dependent upon induction of MAPK suggested that these factors might stably occupy such elements *in vivo*, in contrast to inducible factors which have shorter residency times. These observations suggest that FOX:E motifs may function to maintain an accessible chromatin structure in a natural chromatin context which enables the subsequent binding of inducible factors activated by constitutive FLT3-ITD signalling via MAPK. We also considered the possibility that these inducible factors might bind within the same DHS, but the observation that FOX:E-containing DHS fragments did not feature inducible enhancer activity as tested in

reporter gene assays suggests otherwise. Nevertheless, these data are consistent with a model whereby FOX:E DHSs act as a novel class of non-inducible priming element that somehow cooperate with nearby DHSs to promote gene activation.

In our previous study of T cells, we used CRISPR genome editing to delete a specific pDHS associated with *IL3* expression, directly linking loss of this element to a reduction in target gene expression and validating the priming activity suggested by reporter gene assays. Thus, a logical extension of our current work would be to delete candidate FOX:E motif-containing DHSs and inspect the changes on putative target gene expression. However, our studies in the context of these FOX:E DHSs are still limited by a lack of known cell lines with DHSs containing FOX:E motifs, limiting these functional approaches.

5.8 Conclusions

Using DNA binding assays, we have successfully confirmed that candidate FOX:E composite sites are occupied by nuclear factors *in vitro* in EMSAs. Furthermore, competition assays allowed us to determine that patterns of E-box and FOX motif binding to these composite sites were indeed specific.

Secondly, our luciferase reporter gene assays suggest that DHSs containing FOX:E motifs do not operate as classical enhancers, and did not directly respond to induction of MAPK or calcium signalling following cell stimulation with PMA and CaI. The significance of these findings are not yet clear, but are perhaps consistent with DHSs containing FOX:E elements operating as priming enhancers, as previously described in studies of T cell activation.

5.9 Current limitations and future prospects

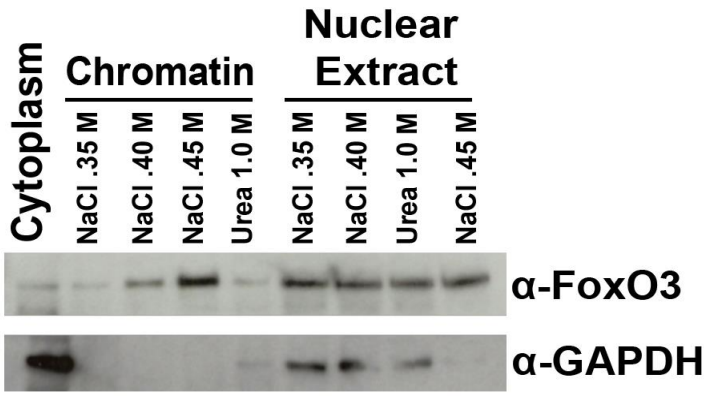
A major limitation of the *in vitro* DNA binding studies to date was that we were unable to confirm the identity of specific factors binding to FOX and E-box motifs in EMSAs. As all of the nuclear extracts we tested in EMSAs featured expression of over 30 FOX proteins and at least as many (if not more) bHLH proteins, to systematically test various antibodies against all of these factors in 'supershift' EMSAs would not be cost-effective.

However, we consider FoxC1 to be the most likely candidate FOX protein binding to these elements *in vivo*, given the highly FLT3-ITD-specific activation of FOX:E elements in parallel with *FOXC1* upregulation. In future experiments, we plan to express recombinant FoxC1 as His-tagged protein in *E. coli* and/or using *in vitro* transcription/translation protocol for use in EMSAs. These studies will allow us to determine the mobility patterns of FoxC1:DNA complexes in EMSAs and greatly increase our confidence that previous results obtained with nuclear extracts are indeed attributable to FOX motif binding by a FOX protein. However, the question of which bHLH factors bind to FOX:E motifs in AML cells still remains.

To address this outstanding question, we could perform double-stranded biotinylated oligonucleotide pulldown assays with AML cell nuclear extracts, and then use mass spectrometry to identify proteins binding to FOX:E motifs in an unbiased manner. These experiments could be used in conjunction with nuclear extract EMSAs and co-immunoprecipitation experiments to shed light on potential *in vivo* protein complexes binding to FOX:E elements, as well as allowing identification of the specific factors binding to probes in EMSAs. These approaches may also identify other potential TFs and/or chromatin regulators that are involved in the activity of FOX:E elements, and would serve as a platform from which to embark on detailed mechanistic studies.

Finally, the question of what the function of FOX:E elements are *in vivo* remains unanswered. To fully address this question, we would need to develop new model systems in which to perform functional analyses, as immortalised AML cell lines investigated thus far, including FUJIOKA cells, lack the DHSs containing FOX:E elements we described in primary AML samples. These could include, for example, the generation of a novel cell line from normal CD34+ PBSCs with overexpression of FLT3-ITD and both HoxA9 with FoxC1. Another alternative would be to perform shRNA knockdown of *FOXC1* in primary AML cells that feature activation of these DHSs. We are already pursuing the latter experiments for other reasons, as described in 4.13, but from the perspective of studying FOX:E motifs, extension of these experiments to include genome-wide profiling of chromatin accessibility and gene expression would allow direct assessment of the effects *FOXC1* depletion has on these DHSs and the potential target genes they are thought to regulate.

Collectively, these additional studies should provide a more detailed analysis of DHSs containing FOX:E sites and their function in AML, and would represent significant mechanistic contributions to the field of gene regulation in leukaemia.



Appendix 5:1 **FoxO3 is expressed in the nuclei of PL-21 cells**

Western blot showing expression of FoxO3 in various subcellular fractions of PL-21 AML cells. A titration of NaCl and Urea buffer conditions were tested to extract histone-associated proteins. GAPDH is used as an internal loading control.

Chapter 6: Overall Discussion

6.1 Indications that FoxC1 may be a pioneer factor

A recurring observation throughout this study was that many DHSs associated with increased *FOXC1* expression lacked an enrichment of canonical FOX binding motifs. These findings were documented both in the context of the Hoxa9+FoxC1-dependent mouse AML model (Chapter 3), as well as in primary human AML samples and cell lines distinguished by high expression of *FOXC1* (Chapter 4). Furthermore, our analyses of FoxC1 binding from two independent ChIP-seq datasets from human and murine tissues suggested that a considerable proportion of FoxC1 binding events occur at nucleosomal sites in the absence of a defined binding motif.

As DHSs are often enriched for TF binding motifs involved in their regulation and function, an absence of a specific TF motif might superficially suggest that this particular factor does not participate directly in the regulation of the DHS[31,419]. This argument is certainly valid for the overwhelming majority of TFs, which generally only recognise target binding sequences as they exist in nucleosome-free chromatin demarcated by DNase I hypersensitivity[137]. However, members of the FOX protein family are distinguished from most other TFs by a common winged-helix DNA-binding domain that mediates sequence-independent binding to compacted chromatin[144,147,148,379]. Indeed, there is compelling evidence indicating that FOX proteins do not just bind to sites in compacted chromatin via a canonical 5'-RYMAAYA-3' FOX motif[140,145].

These observations of both sequence-specific DNA binding and 'non-specific' nucleosome binding by FoxC1 raise questions regarding the functionality of this TF *in vivo*, chiefly that these findings may be consistent with pioneer factor activity. Indeed, an

elegant study from the Zaret laboratory compared binding dynamics of the pioneer factor FoxA1 between asynchronous and mitotic Huh7 liver cells, linking a previously observed bimodal FoxA1 binding pattern to distinct *in vivo* functionality[140], Using fluorescence recovery after photobleaching (FRAP), they demonstrated that FoxA1 was highly mobile in mitotic cells, moving in the nuclei ~2.5 fold more rapidly than in asynchronous cells. Furthermore, by studying FoxA1 mutant proteins lacking the amino acid residues involved in either specific or non-specific modes of binding, they confirmed that this greatly increased nuclear mobility was due to the non-specific binding mode. This effect was observed in FoxA1 ChIP-seq analyses as an increase in non-specific binding enrichment in mitotic cells. Finally, it was determined that the non-specific binding mode enables FoxA1 to act as a mitotic bookmark, priming essential target genes for rapid re-activation following cell division. As the structural determinants conferring bimodal properties of FoxA1 binding occur in the most highly-conserved region of the FoxC1 primary amino acid sequence (Figure 6:6:1), we postulate that FoxC1 may act in a similar manner to this more extensively characterised pioneer factor. A similar series of experiments to those just described by the Zaret laboratory for FoxA1, instead investigating FoxC1, would shed light on the significance of this potential mechanism.

Focusing on the specific mode of binding, recent studies using MNase-seq demonstrated that FoxA proteins often bind nucleosomal FOX motifs at liver-specific enhancers without eviction of the underlying nucleosome[379]. Critically, these nucleosomal sites are not efficiently digested by DNase I enzyme, which preferentially cleaves naked DNA in nucleosome-free regions[2]. Furthermore, another study of FoxA1 binding co-operatively with GR also noted a conspicuous lack of occupied FOX sites at nucleosome-free regions using similar high-resolution DNase I footprinting approaches to our own at high-confidence FoxA1 binding sites identified from ChIP-seq analyses[420]. These

observations raise the possibility that our DNase I footprinting method may simply not detect many FoxC1 binding events, which may not have a sufficiently long residency time on nucleosome-free DNA to leave conspicuous genomic footprints.

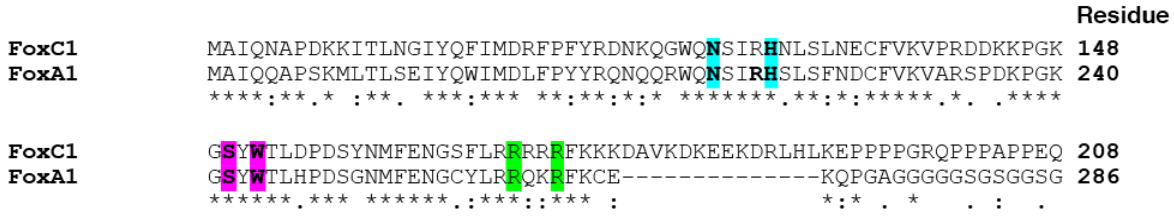


Figure 6:6:1: **Specific residues involved in bimodal binding activity of FoxA1 are conserved in FoxC1.**

Multiple protein sequence alignment of FoxA1 and FoxC1 generated using ClustalW2 (EMBL-EBI) of a portion of the Forkhead DNA-binding domain sequences in FoxC1 and FoxA1. Conserved residues revealed as critical to different binding properties of FoxA1[145] are highlighted.

Blue - Asp and His residues that are essential to normal non-specific binding activity of FoxA1.

Pink - Ser and Trp residues involved in sequence-specific recognition of the canonical FOX motif.

Green - Arg residues essential to both modes of binding.

Alternatively, it is possible that FoxC1 may instead promote the binding of another TF to target DHSs without directly interacting with the DNA itself. Indeed, in studies of estrogen receptor (ER)-positive breast cancer cells, almost all detectable ER binding events were found to depend upon the ectopic overexpression of FoxA1, yet only ~50% of ER binding events overlap directly with FoxA1 binding[401]. These observations raised two possibilities. Firstly, FoxA1 could act as a 'hit and run' factor, using its pioneer activity to open up chromatin for the co-operative binding of further TFs and then dissociating from DNA. However, a more compelling scenario is that FoxA1 instead promoted activation of DHSs by modulating the local chromatin structure. Indeed, a more recent study of ER-positive breast cancer cells demonstrated an association between FoxA1 binding and that of the BRD4, a chromatin factor involved in the nucleation of protein complexes responsible for maintaining active chromatin domains by engaging with acetylated histones, and previously implicated in carcinogenesis[421,422]. Given the considerable

homology between FoxA1 and FoxC1 in key functional domains, we speculate that similar mechanisms may account for our own analyses of FoxC1 binding in AML cells.

6.2 Overcoming the technical limitations of FoxC1 ChIP-seq

Although other methods can offer informative insights into potential regions bound by FoxC1, alone it is insufficient to formally confirm FoxC1 as the TF binding to loci of interest. For this, optimal ChIP conditions need to be established, or alternative methods pursued. We discussed these problems at length with several other groups studying FoxC1, including the laboratories of Tim Somervaille (CRUK Manchester Institute), Elaine Fuchs (Rockefeller University), Rui Yi (University of Colorado, Boulder), Emily Bernstein (Mount Sinai, New York) and Nicoletta Bobola (University of Manchester, whose group generated the embryonic dataset used in our mouse AML analyses in Chapter 3). These discussions revealed that FoxC1 seems to be an especially challenging target to study in conventional ChIP-PCR assays, let alone ChIP-seq. In future experiments, we plan to test a recently released FoxC1 antibody that is being marketed as ChIP-grade, and we have entered discussions with Proteintech who have agreed to send us pre-bleeds of their latest FoxC1 antibody to test in ChIP assays. In parallel, our collaborators at the CRUK Manchester Institute are generating a custom antibody which we will be given access to if they have success. An alternative would be to perform experiments using an epitope-tagged FoxC1 expression construct. Although this would preclude study of primary AML samples, it may at least permit more robust interrogation of FoxC1 activity in appropriate model systems.

Beyond the limitations of ChIP-seq, emerging alternative approaches may permit more detailed interrogation of FoxC1 activity. Firstly, the Henikoff laboratory recently generated a novel approach to interrogate specific TF binding, called 'Cleavage Under Targets and Release Using Nuclease' (CUT&RUN)[423]. Although this method is still reliant on an

antibody, which is used to target a nuclease to sites occupied by the TF of interest, it is a cross-linking free approach performed on intact cell nuclei. The lack of a cross-linking step reduces the potential for the antibody not to recognise TFs of interest by epitope masking, which may also be a limitation of ChIP assays in the study of FoxC1 binding.

Another technique that could be used is the 'CRISPR affinity purification *in situ* of regulatory elements' (CAPTURE) method[424]. Briefly, this approach uses sgRNAs and an inactivated variant of Cas9 to target loci of interest and promote biotinylation of molecules associated with that loci, including TFs, by a separate biotin ligase recruited to loci by the sgRNA/Cas9 complex. Biotinylated macromolecules can then be purified by high-affinity streptavidin pulldown and interrogated in an unbiased approach using mass spectrometry (MS)-based proteomics. This is an appealing alternative to ChIP assays, especially given the high sensitivity of current proteomic approaches to interrogate complex macromolecular complexes, and would permit formal identification of FoxC1 at loci of interest, together with potential interacting partners which could inform future work. An additional benefit of this method is the capability to identify long-range chromatin interactions using 3C-based methods, which would allow the study of FoxC1 in promoting long range 3D chromosomal interactions in AML cells.

6.3 FoxC1: co-opter or co-opted?

Previous studies by Somerville *et al.* (2015) elegantly demonstrated that the overexpression of Hoxa9 and FoxC1 was sufficient to induce an aggressive, transplantable AML in mice, suggesting a functional synergy between these two TFs. Indeed, our analyses performed on Hoxa9+FoxC1 cells reinforced the notion that FoxC1 functionally co-operates with HOX family proteins to reprogram haematopoietic cells during leukaemogenesis, although the specific mechanism by which this occurs remains as yet unclear. We described experiments to address this question in 3.23.

Beyond the evidence favouring a co-operation between HOX proteins and FoxC1 in AML, our analyses throughout these studies also implicated FoxC1 in the re-targeting or disruption of normal haematopoietic TFs. For example, in Hoxa9+FoxC1 AML cells, our integrative accessible chromatin, DNase I footprinting and gene expression analyses all implicated FoxC1 in the redirection of normal Myb and C/EBP activity (Chapter 3). Independently, we observed in ChIP-seq analyses that FoxC1 binding frequently co-associated with C/EBPα and RUNX1 binding in FUJIOKA-specific DHSs (Chapter 4). These data corroborated earlier analyses of primary human AML samples which demonstrated a FLT3-ITD-dependent upregulation of *FOXC1* in parallel with disruption of the normal RUNX1 and C/EBP transcriptional network[271]. Finally, we cannot overlook a stated aim of this project, which was to further understand the role that FOX:E composite sites, highly specific to FLT3-ITD AMLs in primary human samples, but also since found to be a part of the Hoxa9+FoxC1 mouse AML chromatin signature (Chapter 5).

The prospect of FoxC1 co-opting normal haematopoietic TF function is perhaps not surprising, given previously documented roles for other FOX proteins in the oncogenic retargeting of tissue-specific factors. Indeed, this concept was elegantly illustrated in work by the Brown laboratory, which demonstrated cell-type specific collaboration of the pioneer factor FoxA1 with the ER and androgen receptor (AR) in breast and prostate cancers, respectively[78]. Indeed, further studies of ER-positive breast cancer cells revealed that ER binding was almost exclusively dependent upon FoxA1, where FoxA1 promoted ER-mediated transcription of oncogenic target genes[401]. Furthermore, ectopic expression of FoxA1 in either ovarian cancer cells or osteosarcoma cells, which express inactive ER, was sufficient to promote ER binding to known target genes, upregulating the ER-dependent gene expression programme. Finally, of direct relevance to our own studies of FoxC1 in AML, FoxA1-dependent recruitment of ER to target loci occurred

both at nucleosome-free regions and within condensed chromatin, as revealed by accessible chromatin analyses.

In the context of prostate cancer, it was recently demonstrated that forced expression of FoxA1 with HoxB13 in an immortalised prostate cell line caused extensive redistribution of AR binding, causing cells to acquire an epigenetic and gene expression signature consistent with those of primary prostate tumour samples[402]. Independent studies confirmed that AR-target loci in prostate cancer cells were occupied by a complex containing the AR, FoxA1 and HOXB13. Interestingly, in castration-resistant prostate cancer (CRPC), FoxA1 seemed to re-target binding of AR instead to non-canonical AR binding sites, where it may contribute to relapse by the activation of genes involved in invasion and metastasis.

In summary, common roles for FoxA1 overexpression have emerged in distinct cancer types, whereby FoxA1 co-operates with pre-existing tissue-specific TFs to drive epigenetic changes promoting transformation. Similar roles implicating other FOX proteins as oncogenic partners of tissue-specific TFs have been postulated for the FoxO and FoxM1 family members[425,426], increasing confidence that FoxC1 may behave in a similar manner. It is compelling to consider that in the context of developing blood cells, ectopic upregulation of FoxC1 may similarly act to corrupt the activity of existing haematopoietic TFs, contributing to the epigenetic and transcriptional reprogramming that occurs during leukaemogenesis.

6.4 Final conclusions

To our knowledge, we have performed the first integrative genome-wide analysis of chromatin elements and gene expression as related to ectopic expression of FoxC1, not just in AML, but indeed in any cancer. Although our analyses are not yet conclusive, we

have postulated from several independent lines of investigation that FoxC1 may represent a novel pioneer factor which may cause the mistargeting of tissue-specific TFs. This is of potential relevance to several other cancers where roles for FoxC1 have emerged. These include basal-like breast cancer (BLBC), where a phenotypic co-operativity between FoxC1 and the inflammatory TF NFκB has emerged[427]. Here, FoxC1 seems to promote ectopic proliferation, self-renewal and the epithelial-to-mesenchymal transition, although the genome-wide events accompanying these changes have yet to be described[306]. Furthermore, FoxC1 overexpression is a poor prognostic marker in tumours of the liver, stomach, pancreas and endometrium, and in melanoma[302–304,428,429].

The ground has now been laid for a more detailed mechanistic dissection of FoxC1 function, not just in the study of leukaemogenesis, but also for other solid malignancies, where genome-wide characterisation of FoxC1 function is now needed. Critically, the study of TFs and the mechanisms through which they act to drive oncogenic chromatin and gene expression patterns are essential for an improved understanding of the molecular basis of disease, by identifying novel sensitivities that may be targeted in the clinic. Furthermore, despite the longstanding dogma that TFs themselves are 'undruggable' targets in the context of disease therapy, pioneering work led by the Bushweller group and others strongly suggests otherwise[430]. Given that FoxC1 expression is not a feature of normal haematopoietic cells, we propose that FoxC1 may itself be an attractive target for therapy in AML[431]. Finally, given that *FOXC1* expression is associated with subtypes of AML that are refractory to current therapeutic approaches, including FLT3-ITD AMLs[271], identifying novel treatments to improve outcomes in the clinic are needed more than ever.

Bibliography

1. Flemming, W. *Zellsubstanz, Kern und Zelltheilung.* (Leipzig, F. C. W. Vogel, 1882).
2. Cockerill, P. N. Structure and function of active chromatin and DNase I hypersensitive sites. *FEBS J.* **278,** 2182–2210 (2011).
3. Beagrie, R. A. & Pombo, A. Cell cycle: Continuous chromatin changes. *Nature* **547,** 34–35 (2017).
4. Kornberg, R. D. Chromatin structure: a repeating unit of histones and DNA. *Science* **184,** 868–871 (1974).
5. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389,** 251–260 (1997).
6. Arents, G., Burlingame, R. W., Wang, B. C., Love, W. E. & Moudrianakis, E. N. The nucleosomal core histone octamer at 3.1 A resolution: a tripartite protein assembly and a left-handed superhelix. *Proc. Natl. Acad. Sci. U. S. A.* **88,** 10148–10152 (1991).
7. Olins, A. L. & Olins, D. E. Spheroid Chromatin Units (v Bodies). *Science* **183,** 330–332 (1974).
8. Thoma, F., Koller, T. & Klug, A. Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. *J. Cell Biol.* **83,** 403–427 (1979).
9. Hu, Y., Kireev, I., Plutz, M., Ashourian, N. & Belmont, A. S. Large-scale chromatin structure of inducible genes: transcription on a condensed, linear template. *J. Cell Biol.* **185,** 87–100 (2009).
10. Simpson, R. T. Structure of the chromatosome, a chromatin particle containing 160 base pairs of DNA and all the histones. *Biochemistry (Mosc.)* **17,** 5524–5531 (1978).
11. Routh, A., Sandin, S. & Rhodes, D. Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure. *Proc. Natl. Acad. Sci.* **105,** 8872–8877 (2008).
12. Robinson, P. J. J., Fairall, L., Huynh, V. A. T. & Rhodes, D. EM measurements define the dimensions of the “30-nm” chromatin fiber: Evidence for a compact, interdigitated structure. *Proc. Natl. Acad. Sci.* **103,** 6506–6511 (2006).
13. Finch, J. T. & Klug, A. Solenoidal model for superstructure in chromatin. *Proc. Natl. Acad. Sci. U. S. A.* **73,** 1897–1901 (1976).
14. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).
15. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).
16. Tjio, J. H. & Levan, A. The Chromosome Number of Man. *Hereditas* **42,** 1–6 (1956).
17. Cremer, T. & Cremer, M. Chromosome territories. *Cold Spring Harb. Perspect. Biol.* **2,** a003889 (2010).
18. Boyle, S. *et al.* The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum. Mol. Genet.* **10,** 211–219 (2001).
19. Bickmore, W. A. The spatial organization of the human genome. *Annu. Rev. Genomics Hum. Genet.* **14,** 67–84 (2013).
20. Heitz, E. *Das Heterochromatin der Moose*. (Bornträger, 1928).

21. Speicher, M. R., Gwyn Ballard, S. & Ward, D. C. Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nat. Genet.* **12,** 368–375 (1996).
22. Caspersson, T., Zech, L. & Johansson, C. Differential binding of alkylating fluorochromes in human chromosomes. *Exp. Cell Res.* **60,** 315–319 (1970).
23. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453,** 948–951 (2008).
24. Würtele, H. & Chartrand, P. Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology. *Chromosome Res. Int. J. Mol. Supramol. Evol. Asp. Chromosome Biol.* **14,** 477–495 (2006).
25. Stalder, J. *et al.* Tissue-specific DNA cleavages in the globin chromatin domain introduced by DNAase I. *Cell* **20,** 451–460 (1980).
26. Weintraub, H. & Groudine, M. Chromosomal subunits in active genes have an altered conformation. *Science* **193,** 848–856 (1976).
27. Wood, W. I. & Felsenfeld, G. Chromatin structure of the chicken beta-globin gene region. Sensitivity to DNase I, micrococcal nuclease, and DNase II. *J. Biol. Chem.* **257,** 7730–7736 (1982).
28. Bellard, M., Kuo, M. T., Dretzen, G. & Chambon, P. Differential nuclease sensitivity of the ovalbumin and beta-globin chromatin regions in erythrocytes and oviduct cells of laying hen. *Nucleic Acids Res.* **8,** 2737–2750 (1980).
29. Storb, U., Wilson, R., Selsing, E. & Walfield, A. Rearranged and germline immunoglobulin kappa genes: different states of DNase I sensitivity of constant kappa genes in immunocompetent and nonimmune cells. *Biochemistry (Mosc.)* **20,** 990–996 (1981).
30. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10,** 1213–1218 (2013).
31. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132,** 311–322 (2008).
32. Holliday, R. & Pugh, J. E. DNA modification mechanisms and gene activity during development. *Science* **187,** 226–232 (1975).
33. Johnson, T. B. & Coghill, R. D. RESEARCHES ON PYRIMIDINES. C111. THE DISCOVERY OF 5-METHYL-CYTOSINE IN TUBERCULINIC ACID, THE NUCLEIC ACID OF THE TUBERCLE BACILLUS1. *J. Am. Chem. Soc.* **47,** 2838–2844 (1925).
34. Ichiyanagi, T., Ichiyanagi, K., Miyake, M. & Sasaki, H. Accumulation and loss of asymmetric non-CpG methylation during male germ-cell development. *Nucleic Acids Res.* **41,** 738–745 (2013).
35. Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11,** 204–220 (2010).
36. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16,** 6–21 (2002).
37. Goll, M. G. & Bestor, T. H. Eukaryotic cytosine methyltransferases. *Annu. Rev. Biochem.* **74,** 481–514 (2005).
38. Rose, N. R. & Klose, R. J. Understanding the relationship between DNA methylation and histone lysine methylation. *Biochim. Biophys. Acta* **1839,** 1362–1372 (2014).

39. Rösl, F., Arab, A., Klevenz, B. & zur Hausen, H. The effect of DNA methylation on gene regulation of human papillomaviruses. *J. Gen. Virol.* **74 (Pt 5),** 791–801 (1993).
40. Kaneda, M. *et al.* Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature* **429,** 900–903 (2004).
41. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324,** 929–930 (2009).
42. Richa, R. & Sinha, R. P. Hydroxymethylation of DNA: an epigenetic marker. *EXCLI J.* **13,** 592–610 (2014).
43. Allfrey, V. G., Faulkner, R. & Mirsky, A. E. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **51,** 786–794 (1964).
44. Zhao, Y. & Garcia, B. A. Appendix 2 - Comprehensive catalog of currently documented histone modifications. *Cold Spring Harb. Perspect. Biol.* **7,** a025064 (2015).
45. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
46. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403,** 41–45 (2000).
47. Turner, B. M. Cellular memory and the histone code. *Cell* **111,** 285–291 (2002).
48. Hebbes, T. R., Clayton, A. L., Thorne, A. W. & Crane-Robinson, C. Core histone hyperacetylation co-maps with generalized DNase I sensitivity in the chicken beta-globin chromosomal domain. *EMBO J.* **13,** 1823–1830 (1994).
49. Kuo, M.-H. *et al.* Transcription-linked acetylation by Gcn5p of histones H3 and H4 at specific lysines. *Nature* **383,** 269–272 (1996).
50. Bannister, A. J. & Kouzarides, T. The CBP co-activator is a histone acetyltransferase. *Nature* **384,** 641–643 (1996).
51. Taunton, J., Hassig, C. A. & Schreiber, S. L. A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p. *Science* **272,** 408–411 (1996).
52. Zeng, L. & Zhou, M.-M. Bromodomain: an acetyl-lysine binding domain. *FEBS Lett.* **513,** 124–128 (2002).
53. Ogryzko, V. V., Schiltz, R. L., Russanova, V., Howard, B. H. & Nakatani, Y. The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell* **87,** 953–959 (1996).
54. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.* **40,** 897–903 (2008).
55. Yao, T. P. *et al.* Gene dosage-dependent embryonic development and proliferation defects in mice lacking the transcriptional integrator p300. *Cell* **93,** 361–372 (1998).
56. Tie, F. *et al.* CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. *Dev. Camb. Engl.* **136,** 3131–3141 (2009).
57. Shogren-Knaak, M. *et al.* Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science* **311,** 844–847 (2006).
58. Gupta, A. *et al.* The mammalian ortholog of Drosophila MOF that acetylates histone H4 lysine 16 is essential for embryogenesis and oncogenesis. *Mol. Cell. Biol.* **28,** 397–409 (2008).
59. Dion, M. F., Altschuler, S. J., Wu, L. F. & Rando, O. J. Genomic characterization reveals a simple histone H4 acetylation code. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 5501–5506 (2005).

60. Akhtar, A. & Becker, P. B. Activation of transcription through histone H4 acetylation by MOF, an acetyltransferase essential for dosage compensation in Drosophila. *Mol. Cell* **5,** 367–375 (2000).
61. Tschiersch, B. *et al.* The protein encoded by the Drosophila position-effect variegation suppressor gene Su(var)3-9 combines domains of antagonistic regulators of homeotic gene complexes. *EMBO J.* **13,** 3822–3831 (1994).
62. Schotta, G. *et al.* Central role of Drosophila SU(VAR)3-9 in histone H3-K9 methylation and heterochromatic gene silencing. *EMBO J.* **21,** 1121–1131 (2002).
63. Aagaard, L. *et al.* Functional mammalian homologues of the Drosophila PEV-modifier Su(var)3-9 encode centromere-associated proteins which complex with the heterochromatin component M31. *EMBO J.* **18,** 1923–1938 (1999).
64. Singh, P. B. *et al.* A sequence motif found in a Drosophila heterochromatin protein is conserved in animals and plants. *Nucleic Acids Res.* **19,** 789–794 (1991).
65. Cheng, X. Structural and functional coordination of DNA and histone methylation. *Cold Spring Harb. Perspect. Biol.* **6,** (2014).
66. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* **17,** 487–500 (2016).
67. Schraets, D., Lehmann, T., Dingermann, T. & Marschalek, R. MLL-mediated transcriptional gene regulation investigated by gene expression profiling. *Oncogene* **22,** 3655–3668 (2003).
68. Yu, B. D., Hess, J. L., Horning, S. E., Brown, G. A. & Korsmeyer, S. J. Altered Hox expression and segmental identity in Mll-mutant mice. *Nature* **378,** 505–508 (1995).
69. Greer, E. L. & Shi, Y. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat. Rev. Genet.* **13,** 343–357 (2012).
70. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–837 (2007).
71. Chen, K. *et al.* Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. *Nat. Genet.* **47,** 1149–1157 (2015).
72. Pray-Grant, M. G., Daniel, J. A., Schieltz, D., Yates, J. R. & Grant, P. A. Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation. *Nature* **433,** 434–438 (2005).
73. Sims, R. J. *et al.* Human but not yeast CHD1 binds directly and selectively to histone H3 methylated at lysine 4 via its tandem chromodomains. *J. Biol. Chem.* **280,** 41789–41792 (2005).
74. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125,** 315–326 (2006).
75. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470,** 279–283 (2011).
76. Jeong, K. W. *et al.* Recognition of enhancer element-specific histone methylation by TIP60 in transcriptional activation. *Nat. Struct. Mol. Biol.* **18,** 1358–1365 (2011).
77. Orford, K. *et al.* Differential H3K4 methylation identifies developmentally poised hematopoietic genes. *Dev. Cell* **14,** 798–809 (2008).
78. Lupien, M. *et al.* FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132,** 958–970 (2008).

79. Okabe, Y. & Medzhitov, R. Tissue-specific signals control reversible program of localization and functional polarization of macrophages. *Cell* **157,** 832–844 (2014).
80. Sanulli, S. *et al.* Jarid2 Methylation via the PRC2 Complex Regulates H3K27me3 Deposition during Cell Differentiation. *Mol. Cell* **57,** 769–783 (2015).
81. Bartke, T. *et al.* Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell* **143,** 470–484 (2010).
82. Francis, N. J., Kingston, R. E. & Woodcock, C. L. Chromatin compaction by a polycomb group protein complex. *Science* **306,** 1574–1577 (2004).
83. Kadoch, C., Copeland, R. A. & Keilhack, H. PRC2 and SWI/SNF Chromatin Remodeling Complexes in Health and Disease. *Biochemistry (Mosc.)* **55,** 1600–1614 (2016).
84. Kia, S. K., Gorski, M. M., Giannakopoulos, S. & Verrijzer, C. P. SWI/SNF mediates polycomb eviction and epigenetic reprogramming of the INK4b-ARF-INK4a locus. *Mol. Cell. Biol.* **28,** 3457–3464 (2008).
85. Wilson, B. G. *et al.* Epigenetic antagonism between polycomb and SWI/SNF complexes during oncogenic transformation. *Cancer Cell* **18,** 316–328 (2010).
86. Talbert, P. B. & Henikoff, S. Histone variants on the move: substrates for chromatin dynamics. *Nat. Rev. Mol. Cell Biol.* **18,** 115–126 (2017).
87. Thakar, A. *et al.* H2A.Z and H3.3 histone variants affect nucleosome structure: biochemical and biophysical studies. *Biochemistry (Mosc.)* **48,** 10852–10857 (2009).
88. Jin, C. & Felsenfeld, G. Nucleosome stability mediated by histone variants H3.3 and H2A.Z. *Genes Dev.* **21,** 1519–1529 (2007).
89. Chakravarthy, S. & Luger, K. The histone variant macro-H2A preferentially forms 'hybrid nucleosomes'. *J. Biol. Chem.* **281,** 25522–25531 (2006).
90. Volle, C. & Dalal, Y. Histone variants: the tricksters of the chromatin world☆. *Curr. Opin. Genet. Dev.* **0,** 8–14 (2014).
91. Côté, J., Quinn, J., Workman, J. L. & Peterson, C. L. Stimulation of GAL4 derivative binding to nucleosomal DNA by the yeast SWI/SNF complex. *Science* **265,** 53–60 (1994).
92. Kadam, S. & Emerson, B. M. Transcriptional specificity of human SWI/SNF BRG1 and BRM chromatin remodeling complexes. *Mol. Cell* **11,** 377–389 (2003).
93. Weber, C. M., Henikoff, J. G. & Henikoff, S. H2A.Z nucleosomes enriched over active genes are homotypic. *Nat. Struct. Mol. Biol.* **17,** 1500–1507 (2010).
94. Sainsbury, S., Bernecky, C. & Cramer, P. Structural basis of transcription initiation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16,** 129–143 (2015).
95. Grünberg, S. & Hahn, S. Structural insights into transcription initiation by RNA polymerase II. *Trends Biochem. Sci.* **38,** 603–611 (2013).
96. Phatnani, H. P. & Greenleaf, A. L. Phosphorylation and functions of the RNA polymerase II CTD. *Genes Dev.* **20,** 2922–2936 (2006).
97. Weber, C. M., Ramachandran, S. & Henikoff, S. Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Mol. Cell* **53,** 819–830 (2014).
98. Ring, B. Z., Yarnell, W. S. & Roberts, J. W. Function of E. coli RNA polymerase sigma factor sigma 70 in promoter-proximal pausing. *Cell* **86,** 485–493 (1996).

99. Soutoglou, E. & Talianidis, I. Coordination of PIC assembly and chromatin remodeling during differentiation-induced gene activation. *Science* **295,** 1901–1904 (2002).
100. Krebs, A. R. *et al.* Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Mol. Cell* doi:10.1016/j.molcel.2017.06.027
101. Peterlin, B. M. & Price, D. H. Controlling the elongation phase of transcription with P-TEFb. *Mol. Cell* **23,** 297–305 (2006).
102. Chen, F. X. *et al.* PAF1, a Molecular Regulator of Promoter-Proximal Pausing by RNA Polymerase II. *Cell* **162,** 1003–1015 (2015).
103. Belotserkovskaya, R. *et al.* FACT facilitates transcription-dependent nucleosome alteration. *Science* **301,** 1090–1093 (2003).
104. Orphanides, G., Wu, W. H., Lane, W. S., Hampsey, M. & Reinberg, D. The chromatin-specific transcription elongation factor FACT comprises human SPT16 and SSRP1 proteins. *Nature* **400,** 284–288 (1999).
105. Jonkers, I., Kwak, H. & Lis, J. T. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife* **3,** e02407 (2014).
106. Skourti-Stathaki, K., Kamieniarz-Gdula, K. & Proudfoot, N. J. R-loops induce repressive chromatin marks over mammalian gene terminators. *Nature* **516,** 436–439 (2014).
107. Skourti-Stathaki, K., Proudfoot, N. J. & Gromak, N. Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. *Mol. Cell* **42,** 794–805 (2011).
108. Porrua, O. & Libri, D. Transcription termination and the control of the transcriptome: why, where and how to stop. *Nat. Rev. Mol. Cell Biol.* **16,** 190–202 (2015).
109. Larson, D. R. *et al.* Direct observation of frequency modulated transcription in single cells using light activation. *eLife* **2,** e00750 (2013).
110. Yudkovsky, N., Ranish, J. A. & Hahn, S. A transcription reinitiation intermediate that is stabilized by activator. *Nature* **408,** 225–229 (2000).
111. Jackson, D. A., Hassan, A. B., Errington, R. J. & Cook, P. R. Visualization of focal sites of transcription within human nuclei. *EMBO J.* **12,** 1059–1065 (1993).
112. Martin, S. & Pombo, A. Transcription factories: quantitative studies of nanostructures in the mammalian nucleus. *Chromosome Res. Int. J. Mol. Supramol. Evol. Asp. Chromosome Biol.* **11,** 461–470 (2003).
113. Papantonis, A. & Cook, P. R. Transcription factories: genome organization and gene regulation. *Chem. Rev.* **113,** 8683–8705 (2013).
114. Waddington, C. H. Canalization of Development and the Inheritance of Acquired Characters. *Nature* **150,** 563–565 (1942).
115. Waddington, C. H. *The strategy of the genes : a discussion of some aspects of theoretical biology*. (London : Allen & Unwin, 1957).
116. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3,** 318–356 (1961).
117. Lewis, E. B. A gene complex controlling segmentation in Drosophila. *Nature* **276,** 565–570 (1978).
118. Quinonez, S. C. & Innis, J. W. Human HOX gene disorders. *Mol. Genet. Metab.* **111,** 4–15 (2014).
119. McGinnis, W. & Krumlauf, R. Homeobox genes and axial patterning. *Cell* **68,** 283–302 (1992).

120. Davis, R. L., Weintraub, H. & Lassar, A. B. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* **51,** 987–1000 (1987).
121. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126,** 663–676 (2006).
122. Jackson, B. C., Carpenter, C., Nebert, D. W. & Vasiliou, V. Update of human and mouse forkhead box (FOX) gene families. *Hum. Genomics* **4,** 345–352 (2010).
123. Ptashne, M. & Gann, A. Transcriptional activation by recruitment. *Nature* **386,** 569–577 (1997).
124. Jonkers, I. & Lis, J. T. Getting up to speed with transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16,** 167–177 (2015).
125. Orphanides, G., Lagrange, T. & Reinberg, D. The general transcription factors of RNA polymerase II. *Genes Dev.* **10,** 2657–2683 (1996).
126. de la Serna, I. L. *et al.* MyoD targets chromatin remodeling complexes to the myogenin locus prior to forming a stable DNA-bound complex. *Mol. Cell. Biol.* **25,** 3997–4009 (2005).
127. Levantini, E. *et al.* RUNX1 regulates the CD34 gene in haematopoietic stem cells by mediating interactions with a distal regulatory element. *EMBO J.* **30,** 4059–4070 (2011).
128. Chen, Y. *et al.* DNA Binding by GATA Transcription Factor Suggests Mechanisms of DNA Looping and Long-Range Gene Regulation. *Cell Rep.* **2,** 1197–1206 (2012).
129. Amoutzias, G. D. *et al.* One billion years of bZIP transcription factor evolution: conservation and change in dimerization and DNA-binding site specificity. *Mol. Biol. Evol.* **24,** 827–835 (2007).
130. Sauer, F. & Tjian, R. Mechanisms of transcriptional activation: differences and similarities between yeast, Drosophila, and man. *Curr. Opin. Genet. Dev.* **7,** 176–181 (1997).
131. Goodman, R. H. & Smolik, S. CBP/p300 in cell growth, transformation, and development. *Genes Dev.* **14,** 1553–1577 (2000).
132. Ng, K. W., Ridgway, P., Cohen, D. R. & Tremethick, D. J. The binding of a Fos/Jun heterodimer can completely disrupt the structure of a nucleosome. *EMBO J.* **16,** 2072–2085 (1997).
133. Adams, C. C. & Workman, J. L. Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. *Mol. Cell. Biol.* **15,** 1405–1421 (1995).
134. Chronis, C. *et al.* Cooperative Binding of Transcription Factors Orchestrates Reprogramming. *Cell* **168,** 442–459.e20 (2017).
135. Voss, T. C. *et al.* Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism. *Cell* **146,** 544–554 (2011).
136. Schalch, T., Duda, S., Sargent, D. F. & Richmond, T. J. X-ray structure of a tetranucleosome and its implications for the chromatin fibre. *Nature* **436,** 138–141 (2005).
137. Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.* **25,** 2227–2241 (2011).
138. Taslim, C. *et al.* Integrated analysis identifies a class of androgen-responsive genes regulated by short combinatorial long-range mechanism facilitated by CTCF. *Nucleic Acids Res.* **40,** 4754–4764 (2012).
139. Carroll, J. S. *et al.* Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* **122,** 33–43 (2005).
140. Caravaca, J. M. *et al.* Bookmarking by specific and nonspecific binding of FoxA1 pioneer factor to mitotic chromosomes. *Genes Dev.* **27,** 251–260 (2013).

141. McPherson, C. E., Shim, E. Y., Friedman, D. S. & Zaret, K. S. An active tissue-specific enhancer and bound transcription factors existing in a precisely positioned nucleosomal array. *Cell* **75,** 387–398 (1993).
142. Cirillo, L. A. & Zaret, K. S. An early developmental transcription factor complex that is more stable on nucleosome core particles than on free DNA. *Mol. Cell* **4,** 961–969 (1999).
143. 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).
144. Cirillo, L. A. *et al.* Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome. *EMBO J.* **17,** 244–254 (1998).
145. Sekiya, T., Muthurajan, U. M., Luger, K., Tulin, A. V. & Zaret, K. S. Nucleosome-binding affinity as a primary determinant of the nuclear mobility of the pioneer transcription factor FoxA. *Genes Dev.* **23,** 804–809 (2009).
146. Lam, E. W.-F., Brosens, J. J., Gomes, A. R. & Koo, C.-Y. Forkhead box proteins: tuning forks for transcriptional harmony. *Nat. Rev. Cancer* **13,** 482–495 (2013).
147. Hatta, M. & Cirillo, L. A. Chromatin opening and stable perturbation of core histone:DNA contacts by FoxO1. *J. Biol. Chem.* **282,** 35583–35593 (2007).
148. Cuesta, I., Zaret, K. S. & Santisteban, P. The forkhead factor FoxE1 binds to the thyroperoxidase promoter during thyroid cell differentiation and modifies compacted chromatin structure. *Mol. Cell. Biol.* **27,** 7302–7314 (2007).
149. Soufi, A. *et al.* Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161,** 555–568 (2015).
150. Zaret, K. S., Lerner, J. & Iwafuchi-Doi, M. Chromatin Scanning by Dynamic Binding of Pioneer Factors. *Mol. Cell* **62,** 665–667 (2016).
151. Consortium, T. E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
152. Ponting, C. P. & Hardison, R. C. What fraction of the human genome is functional? *Genome Res.* **21,** 1769–1776 (2011).
153. Dekker, J. & Misteli, T. Long-Range Chromatin Interactions. *Cold Spring Harb. Perspect. Biol.* **7,** a019356 (2015).
154. Kadonaga, J. T. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip. Rev. Dev. Biol.* **1,** 40–51 (2012).
155. Lifton, R. P., Goldberg, M. L., Karp, R. W. & Hogness, D. S. The Organization of the Histone Genes in Drosophila melanogaster: Functional and Evolutionary Implications. *Cold Spring Harb. Symp. Quant. Biol.* **42,** 1047–1051 (1978).
156. Javahery, R., Khachi, A., Lo, K., Zenzie-Gregory, B. & Smale, S. T. DNA sequence requirements for transcriptional initiator activity in mammalian cells. *Mol. Cell. Biol.* **14,** 116–127 (1994).
157. Smale, S. T. & Baltimore, D. The 'initiator' as a transcription control element. *Cell* **57,** 103–113 (1989).
158. Kutach, A. K. & Kadonaga, J. T. The Downstream Promoter Element DPE Appears To Be as Widely Used as the TATA Box in Drosophila Core Promoters. *Mol. Cell. Biol.* **20,** 4754–4764 (2000).
159. Bucher, P. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* **212,** 563–578 (1990).

160. Blake, M. C., Jambou, R. C., Swick, A. G., Kahn, J. W. & Azizkhan, J. C. Transcriptional initiation is controlled by upstream GC-box interactions in a TATAA-less promoter. *Mol. Cell. Biol.* **10,** 6632–6641 (1990).
161. Suske, G., Bruford, E. & Philipsen, S. Mammalian SP/KLF transcription factors: bring in the family. *Genomics* **85,** 551–556 (2005).
162. Dynan, W. S. & Tjian, R. The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. *Cell* **35,** 79–87 (1983).
163. Efstratiadis, A. *et al.* The structure and evolution of the human beta-globin gene family. *Cell* **21,** 653–668 (1980).
164. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 1412–1417 (2006).
165. Sur, I. & Taipale, J. The role of enhancers in cancer. *Nat. Rev. Cancer* **16,** 483–493 (2016).
166. Kim, T.-K. & Shiekhattar, R. Architectural and Functional Commonalities between Enhancers and Promoters. *Cell* **162,** 948–959 (2015).
167. Neuberger, M. S. Expression and regulation of immunoglobulin heavy chain gene transfected into lymphoid cells. *EMBO J.* **2,** 1373–1378 (1983).
168. Gillies, S. D., Morrison, S. L., Oi, V. T. & Tonegawa, S. A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell* **33,** 717–728 (1983).
169. Benoist, C. & Chambon, P. Deletions covering the putative promoter region of early mRNAs of simian virus 40 do not abolish T-antigen expression. *Proc. Natl. Acad. Sci. U. S. A.* **77,** 3865–3869 (1980).
170. Grosschedl, R. & Birnstiel, M. L. Spacer DNA sequences upstream of the T-A-T-A-A-A-T-A sequence are essential for promotion of H2A histone gene transcription in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **77,** 7102–7106 (1980).
171. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27,** 299–308 (1981).
172. Banerji, J., Olson, L. & Schaffner, W. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33,** 729–740 (1983).
173. Sharpe, J., Nonchev, S., Gould, A., Whiting, J. & Krumlauf, R. Selectivity, sharing and competitive interactions in the regulation of Hoxb genes. *EMBO J.* **17,** 1788–1798 (1998).
174. Goto, T., Macdonald, P. & Maniatis, T. Early and late periodic patterns of even skipped expression are controlled by distinct regulatory elements that respond to different spatial cues. *Cell* **57,** 413–422 (1989).
175. Grosveld, F., van Assendelft, G. B., Greaves, D. R. & Kollias, G. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell* **51,** 975–985 (1987).
176. Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467,** 430–435 (2010).
177. Hadjur, S. *et al.* Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature* **460,** 410–413 (2009).

178. Ho, Y., Elefant, F., Liebhaber, S. A. & Cooke, N. E. Locus control region transcription plays an active role in long-range gene activation. *Mol. Cell* **23,** 365–375 (2006).
179. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155,** 934–947 (2013).
180. Beagrie, R. A. *et al.* Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543,** 519–524 (2017).
181. Liu, Z. *et al.* 3D imaging of Sox2 enhancer clusters in embryonic stem cells. *eLife* **3,** e04236 (2014).
182. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27,** 38–52 (2017).
183. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* gr.173518.114 (2014). doi:10.1101/gr.173518.114
184. Bevington, S. L. *et al.* Inducible chromatin priming is associated with the establishment of immunological memory in T cells. *EMBO J.* e201592534 (2016). doi:10.15252/embj.201592534
185. Kolovos, P., Knoch, T. A., Grosveld, F. G., Cook, P. R. & Papantonis, A. Enhancers and silencers: an integrated and simple model for their function. *Epigenetics Chromatin* **5,** 1 (2012).
186. Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351,** 1454–1458 (2016).
187. Barkess, G. & West, A. G. Chromatin insulator elements: establishing barriers to set heterochromatin boundaries. *Epigenomics* **4,** 67–80 (2012).
188. Liu, M. *et al.* Genomic discovery of potent chromatin insulators for human gene therapy. *Nat. Biotechnol.* **33,** 198–203 (2015).
189. Hnisz, D., Day, D. S. & Young, R. A. Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell* **167,** 1188–1200 (2016).
190. de Wit, E. *et al.* CTCF Binding Polarity Determines Chromatin Looping. *Mol. Cell* **60,** 676–684 (2015).
191. Ji, X. *et al.* 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. *Cell Stem Cell* **18,** 262–275 (2016).
192. Lefevre, P., Witham, J., Lacroix, C. E., Cockerill, P. N. & Bonifer, C. The LPS-Induced Transcriptional Upregulation of the Chicken Lysozyme Locus Involves CTCF Eviction and Noncoding RNA Transcription. *Mol. Cell* **32,** 129–139 (2008).
193. Baniahmad, A., Muller, M., Steiner, C. & Renkawitz, R. Activity of two different silencer elements of the chicken lysozyme gene can be compensated by enhancer elements. *EMBO J.* **6,** 2297–2303 (1987).
194. Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529,** 110–114 (2016).
195. Koltzenburg, C. Maximow 1909 English version. Available at: http://www.ctt-journal.com/1-3-en-maximow-1909-translation.html.
196. Till, J. E. & McCulloch, E. A. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiat. Res.* **14,** 213–222 (1961).
197. Becker, A. J., McCulloch, E. A. & Till, J. E. Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. *Nature* **197,** 452–454 (1963).
198. Muller-Sieburg, C. E., Whitlock, C. A. & Weissman, I. L. Isolation of two early B lymphocyte progenitors from mouse marrow: a committed pre-pre-B cell and a clonogenic Thy-1-lo hematopoietic stem cell. *Cell* **44,** 653–662 (1986).

199. Spangrude, G. J., Heimfeld, S. & Weissman, I. L. Purification and characterization of mouse hematopoietic stem cells. *Science* **241,** 58–62 (1988).
200. Morrison, S. J. & Spradling, A. C. Stem cells and niches: mechanisms that promote stem cell maintenance throughout life. *Cell* **132,** 598–611 (2008).
201. Morrison, S. J. & Scadden, D. T. The bone marrow niche for haematopoietic stem cells. *Nature* **505,** 327–334 (2014).
202. Nombela-Arrieta, C. *et al.* Quantitative imaging of haematopoietic stem and progenitor cell localization and hypoxic status in the bone marrow microenvironment. *Nat. Cell Biol.* **15,** 533–543 (2013).
203. Testa, U., Labbaye, C., Castelli, G. & Pelosi, E. Oxidative stress and hypoxia in normal and leukemic stem cells. *Exp. Hematol.* **44,** 540–560 (2016).
204. Sacchetti, B. *et al.* Self-renewing osteoprogenitors in bone marrow sinusoids can organize a hematopoietic microenvironment. *Cell* **131,** 324–336 (2007).
205. Hooper, A. T. *et al.* Engraftment and reconstitution of hematopoiesis is dependent on VEGFR2-mediated regeneration of sinusoidal endothelial cells. *Cell Stem Cell* **4,** 263–274 (2009).
206. Katayama, Y. *et al.* Signals from the sympathetic nervous system regulate hematopoietic stem cell egress from bone marrow. *Cell* **124,** 407–421 (2006).
207. Méndez-Ferrer, S., Lucas, D., Battista, M. & Frenette, P. S. Haematopoietic stem cell release is regulated by circadian oscillations. *Nature* **452,** 442–447 (2008).
208. Akashi, K., Traver, D., Miyamoto, T. & Weissman, I. L. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature* **404,** 193–197 (2000).
209. Morrison, S. J., Wandycz, A. M., Hemmati, H. D., Wright, D. E. & Weissman, I. L. Identification of a lineage of multipotent hematopoietic progenitors. *Dev. Camb. Engl.* **124,** 1929–1939 (1997).
210. Orkin, S. H. Diversification of haematopoietic stem cells to specific lineages. *Nat. Rev. Genet.* **1,** 57–64 (2000).
211. Christensen, J. L. & Weissman, I. L. Flk-2 is a marker in hematopoietic stem cell differentiation: a simple method to isolate long-term stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 14541–14546 (2001).
212. Ciofani, M. & Zúñiga-Pflücker, J. C. The thymus as an inductive site for T lymphopoiesis. *Annu. Rev. Cell Dev. Biol.* **23,** 463–493 (2007).
213. Zhang, Y., Garcia-Ibanez, L. & Toellner, K.-M. Regulation of germinal center B-cell differentiation. *Immunol. Rev.* **270,** 8–19 (2016).
214. Adolfsson, J. *et al.* Upregulation of Flt3 expression within the bone marrow Lin(-)Sca1(+)c-kit(+) stem cell compartment is accompanied by loss of self-renewal capacity. *Immunity* **15,** 659–669 (2001).
215. Adolfsson, J. *et al.* Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. *Cell* **121,** 295–306 (2005).
216. Notta, F. *et al.* Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351,** aab2116 (2016).
217. Görgens, A. *et al.* Revision of the human hematopoietic tree: granulocyte subtypes derive from distinct hematopoietic lineages. *Cell Rep.* **3,** 1539–1552 (2013).
218. Yamamoto, R. *et al.* Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells. *Cell* **154,** 1112–1126 (2013).

219. Naik, S. H. *et al.* Diverse and heritable lineage imprinting of early haematopoietic progenitors. *Nature* **496,** 229–232 (2013).
220. Velten, L. *et al.* Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.* **19,** 271–281 (2017).
221. Yu, V. W. C. *et al.* Epigenetic Memory Underlies Cell-Autonomous Heterogeneous Behavior of Hematopoietic Stem Cells. *Cell* **167,** 1310–1322.e17 (2016).
222. Melnikova, I. N., Crute, B. E., Wang, S. & Speck, N. A. Sequence specificity of the core-binding factor. *J. Virol.* **67,** 2408–2411 (1993).
223. North, T. *et al.* Cbfa2 is required for the formation of intra-aortic hematopoietic clusters. *Dev. Camb. Engl.* **126,** 2563–2575 (1999).
224. Lorsbach, R. B. *et al.* Role of RUNX1 in adult hematopoiesis: analysis of RUNX1-IRES-GFP knock-in mice reveals differential lineage expression. *Blood* **103,** 2522–2529 (2004).
225. Lancrin, C. *et al.* The haemangioblast generates haematopoietic cells through a haemogenic endothelium stage. *Nature* **457,** 892–895 (2009).
226. Follows, G. A. *et al.* Epigenetic consequences of AML1-ETO action at the human c-FMS locus. *EMBO J.* **22,** 2798–2809 (2003).
227. Hoogenkamp, M. *et al.* The Pu.1 locus is differentially regulated at the level of chromatin structure and noncoding transcription by alternate mechanisms at distinct developmental stages of hematopoiesis. *Mol. Cell. Biol.* **27,** 7425–7438 (2007).
228. Scott, E. W., Simon, M. C., Anastasi, J. & Singh, H. Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages. *Science* **265,** 1573–1577 (1994).
229. Iwasaki, H. *et al.* Distinctive and indispensable roles of PU.1 in maintenance of hematopoietic stem cells and their differentiation. *Blood* **106,** 1590–1600 (2005).
230. Tenen, D. G., Hromas, R., Licht, J. D. & Zhang, D. E. Transcription factors, normal myeloid development, and leukemia. *Blood* **90,** 489–519 (1997).
231. Zhang, P. *et al.* PU.1 inhibits GATA-1 function and erythroid differentiation by blocking GATA-1 DNA binding. *Blood* **96,** 2641–2648 (2000).
232. Rekhtman, N., Radparvar, F., Evans, T. & Skoultchi, A. I. Direct interaction of hematopoietic transcription factors PU.1 and GATA-1: functional antagonism in erythroid cells. *Genes Dev.* **13,** 1398–1411 (1999).
233. Nerlov, C., Querfurth, E., Kulessa, H. & Graf, T. GATA-1 interacts with the myeloid PU.1 transcription factor and represses PU.1-dependent transcription. *Blood* **95,** 2543–2551 (2000).
234. Chou, S. T. *et al.* Graded repression of PU.1/Sfpi1 gene transcription by GATA factors regulates hematopoietic cell fate. *Blood* **114,** 983–994 (2009).
235. Hoppe, P. S. *et al.* Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. *Nature* **535,** 299–302 (2016).
236. Radomska, H. S. *et al.* CCAAT/enhancer binding protein alpha is a regulatory switch sufficient for induction of granulocytic development from bipotential myeloid progenitors. *Mol. Cell. Biol.* **18,** 4301–4314 (1998).
237. Ramji, D. P. & Foka, P. CCAAT/enhancer-binding proteins: structure, function and regulation. *Biochem. J.* **365,** 561–575 (2002).

238. Zhang, D.-E. *et al.* Absence of granulocyte colony-stimulating factor signaling and neutrophil development in CCAAT enhancer binding protein α-deficient mice. *Proc. Natl. Acad. Sci.* **94,** 569–574 (1997).
239. Zhang, P. *et al.* Enhancement of hematopoietic stem cell repopulating capacity and self-renewal in the absence of the transcription factor C/EBP alpha. *Immunity* **21,** 853–863 (2004).
240. Zhang, D. E. *et al.* CCAAT enhancer-binding protein (C/EBP) and AML1 (CBF alpha2) synergistically activate the macrophage colony-stimulating factor receptor promoter. *Mol. Cell. Biol.* **16,** 1231–1240 (1996).
241. Friedman, A. D. Transcriptional control of granulocyte and monocyte development. *Oncogene* **26,** 6816–6828 (2007).
242. Huang, W., Horvath, E. & Eklund, E. A. PU.1, interferon regulatory factor (IRF) 2, and the interferon consensus sequence-binding protein (ICSBP/IRF8) cooperate to activate NF1 transcription in differentiating myeloid cells. *J. Biol. Chem.* **282,** 6629–6643 (2007).
243. Grondin, B. *et al.* c-Jun homodimers can function as a context-specific coactivator. *Mol. Cell. Biol.* **27,** 2919–2933 (2007).
244. Rangatia, J. *et al.* Elevated c-Jun expression in acute myeloid leukemias inhibits C/EBPalpha DNA binding via leucine zipper domain interaction. *Oncogene* **22,** 4760–4764 (2003).
245. Lidonnici, M. R. *et al.* Expression of the transcriptional repressor Gfi-1 is regulated by C/EBP{alpha} and is involved in its proliferation and colony formation-inhibitory effects in p210BCR/ABL-expressing cells. *Cancer Res.* **70,** 7949–7959 (2010).
246. Dahl, R., Iyer, S. R., Owens, K. S., Cuylear, D. D. & Simon, M. C. The Transcriptional Repressor GFI-1 Antagonizes PU.1 Activity through Protein-Protein Interaction. *J. Biol. Chem.* **282,** 6473–6483 (2007).
247. Yamanaka, R. *et al.* CCAAT/enhancer binding protein ε is preferentially up-regulated during granulocytic differentiation and its functional versatility is determined by alternative use of promoters and differential splicing. *Proc. Natl. Acad. Sci.* **94,** 6462–6467 (1997).
248. Kumar, C. C. Genetic Abnormalities and Challenges in the Treatment of Acute Myeloid Leukemia. *Genes Cancer* **2,** 95–107 (2011).
249. Vardiman, J. W. *et al.* The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood* **114,** 937–951 (2009).
250. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194,** 23–28 (1976).
251. Grove, C. S. & Vassiliou, G. S. Acute myeloid leukaemia: a paradigm for the clonal evolution of cancer? *Dis. Model. Mech.* **7,** 941–951 (2014).
252. Gilliland, D. G. & Griffin, J. D. The roles of FLT3 in hematopoiesis and leukemia. *Blood* **100,** 1532–1542 (2002).
253. Rosnet, O. *et al.* Human FLT3/FLK2 gene: cDNA cloning and expression in hematopoietic cells. *Blood* **82,** 1110–1119 (1993).
254. Ray, R. J., Paige, C. J., Furlonger, C., Lyman, S. D. & Rottapel, R. Flt3 ligand supports the differentiation of early B cell progenitors in the presence of interleukin-11 and interleukin-7. *Eur. J. Immunol.* **26,** 1504–1510 (1996).

255. Rusten, L. S., Lyman, S. D., Veiby, O. P. & Jacobsen, S. E. The FLT3 ligand is a direct and potent stimulator of the growth of primitive and committed human CD34+ bone marrow progenitor cells in vitro. *Blood* **87,** 1317–1325 (1996).
256. Schnittger, S. *et al.* Analysis of FLT3 length mutations in 1003 patients with acute myeloid leukemia: correlation to cytogenetics, FAB subtype, and prognosis in the AMLCG study and usefulness as a marker for the detection of minimal residual disease. *Blood* **100,** 59–66 (2002).
257. Kiyoi, H., Ohno, R., Ueda, R., Saito, H. & Naoe, T. Mechanism of constitutive activation of FLT3 with internal tandem duplication in the juxtamembrane domain. *Oncogene* **21,** 2555–2563 (2002).
258. Shih, A. H., Abdel-Wahab, O., Patel, J. P. & Levine, R. L. The role of mutations in epigenetic regulators in myeloid malignancies. *Nat. Rev. Cancer* **12,** 599–612 (2012).
259. Figueroa, M. E. *et al.* Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. *Cancer Cell* **18,** 553–567 (2010).
260. Patel, J. P. *et al.* Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. *N. Engl. J. Med.* **366,** 1079–1089 (2012).
261. Beerman, I., Seita, J., Inlay, M. A., Weissman, I. L. & Rossi, D. J. Quiescent hematopoietic stem cells accumulate DNA damage during aging that is repaired upon entry into cell cycle. *Cell Stem Cell* **15,** 37–50 (2014).
262. Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat. Med.* **3,** 730–737 (1997).
263. Jan, M. *et al.* Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Sci. Transl. Med.* **4,** 149ra118 (2012).
264. Sykes, S. M., Kokkaliaris, K. D., Milsom, M. D., Levine, R. L. & Majeti, R. Clonal evolution of preleukemic hematopoietic stem cells in acute myeloid leukemia. *Exp. Hematol.* **43,** 989–992 (2015).
265. Shlush, L. I. *et al.* Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nature* **547,** 104–108 (2017).
266. Turhan, A. G. *et al.* Highly purified primitive hematopoietic stem cells are PML-RARA negative and generate nonclonal progenitors in acute promyelocytic leukemia. *Blood* **85,** 2154–2161 (1995).
267. Tallman, M. S., Nabhan, C., Feusner, J. H. & Rowe, J. M. Acute promyelocytic leukemia: evolving therapeutic strategies. *Blood* **99,** 759–767 (2002).
268. Seifert, M. *et al.* Cellular origin and pathophysiology of chronic lymphocytic leukemia. *J. Exp. Med.* **209,** 2183–2198 (2012).
269. Berquam-Vrieze, K. E. *et al.* Cell of origin strongly influences genetic selection in a mouse model of T-ALL. *Blood* **118,** 4646–4656 (2011).
270. Dunphy, C. H., van Deventer, H. W., Carder, K. J., Rao, K. W. & Dent, G. A. Mature B-cell acute lymphoblastic leukemia with associated translocations (14;18)(q32;q21) and (8;9)(q24;p13). A Burkitt variant? *Arch. Pathol. Lab. Med.* **127,** 610–613 (2003).
271. Cauchy, P. *et al.* Chronic FLT3-ITD Signaling in Acute Myeloid Leukemia Is Connected to a Specific Chromatin Signature. *Cell Rep.* **12,** 821–836 (2015).
272. Sood, R., Kamikubo, Y. & Liu, P. Role of RUNX1 in hematological malignancies. *Blood* **129,** 2070–2082 (2017).

273. Gaidzik, V. I. *et al.* RUNX1 mutations in acute myeloid leukemia are associated with distinct clinico-pathologic and genetic features. *Leukemia* **30,** 2282 (2016).
274. Speck, N. A. & Gilliland, D. G. Core-binding factors in haematopoiesis and leukaemia. *Nat. Rev. Cancer* **2,** 502–513 (2002).
275. Miyoshi, H. *et al.* t(8;21) breakpoints on chromosome 21 in acute myeloid leukemia are clustered within a limited region of a single gene, AML1. *Proc. Natl. Acad. Sci.* **88,** 10431–10434 (1991).
276. Wang, J., Hoshino, T., Redner, R. L., Kajigaya, S. & Liu, J. M. ETO, fusion partner in t(8;21) acute myeloid leukemia, represses transcription by interaction with the human N-CoR/mSin3/HDAC1 complex. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 10860–10865 (1998).
277. Pabst, T. *et al.* AML1-ETO downregulates the granulocytic differentiation factor C/EBPalpha in t(8;21) myeloid leukemia. *Nat. Med.* **7,** 444–451 (2001).
278. Kuchenbauer, F., Feuring-Buske, M. & Buske, C. AML1-ETO needs a partner: new insights into the pathogenesis of t(8;21) leukemia. *Cell Cycle Georget. Tex* **4,** 1716–1718 (2005).
279. Ptasinska, A. *et al.* Identification of a dynamic core transcriptional network in t(8;21) AML that regulates differentiation block and self-renewal. *Cell Rep.* **8,** 1974–1988 (2014).
280. Sun, X.-J. *et al.* A stable transcription factor complex nucleated by oligomeric AML1-ETO controls leukaemogenesis. *Nature* **500,** 93–97 (2013).
281. Vangala, R. K. *et al.* The myeloid master regulator transcription factor PU.1 is inactivated by AML1-ETO in t(8;21) myeloid leukemia. *Blood* **101,** 270–277 (2003).
282. Ptasinska, A. *et al.* Depletion of RUNX1/ETO in t(8;21) AML cells leads to genome-wide changes in chromatin structure and transcription factor binding. *Leukemia* **26,** 1829–1841 (2012).
283. Nucifora, G. *et al.* Consistent intergenic splicing and production of multiple transcripts between AML1 at 21q22 and unrelated genes at 3q26 in (3;21)(q26;q22) translocations. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 4004–4008 (1994).
284. Morishita, K., Suzukawa, K., Taki, T., Ihle, J. N. & Yokota, J. EVI-1 zinc finger protein works as a transcriptional activator via binding to a consensus sequence of GACAAGATAAGATAAN1-28 CTCATCTTC. *Oncogene* **10,** 1961–1967 (1995).
285. Goyama, S. *et al.* Evi-1 is a critical regulator for hematopoietic stem cells and transformed leukemic cells. *Cell Stem Cell* **3,** 207–220 (2008).
286. Castilla, L. H. *et al.* Failure of embryonic hematopoiesis and lethal hemorrhages in mouse embryos heterozygous for a knocked-in leukemia gene CBFB-MYH11. *Cell* **87,** 687–696 (1996).
287. Wang, Q. *et al.* The CBFbeta subunit is essential for CBFalpha2 (AML1) function in vivo. *Cell* **87,** 697–708 (1996).
288. Liu, Y. *et al.* Structural basis for recognition of SMRT/N-CoR by the MYND domain and its contribution to AML1/ETO's activity. *Cancer Cell* **11,** 483–497 (2007).
289. Osato, M. *et al.* Biallelic and heterozygous point mutations in the runt domain of the AML1/PEBP2alphaB gene associated with myeloblastic leukemias. *Blood* **93,** 1817–1824 (1999).
290. Wouters, B. J. *et al.* Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome. *Blood* **113,** 3088–3091 (2009).

291. Lin, F. T., MacDougald, O. A., Diehl, A. M. & Lane, M. D. A 30-kDa alternative translation product of the CCAAT/enhancer binding protein alpha message: transcriptional activator lacking antimitotic activity. *Proc. Natl. Acad. Sci. U. S. A.* **90,** 9606–9610 (1993).
292. Kirstetter, P. *et al.* Modeling of C/EBPalpha mutant acute myeloid leukemia reveals a common expression signature of committed myeloid leukemia-initiating cells. *Cancer Cell* **13,** 299–310 (2008).
293. Schwieger, M. *et al.* A dominant-negative mutant of C/EBPalpha, associated with acute myeloid leukemias, inhibits differentiation of myeloid and erythroid progenitors of man but not mouse. *Blood* **103,** 2744–2752 (2004).
294. Hughes, J. M. *et al.* C/EBPα-p30 protein induces expression of the oncogenic long non-coding RNA UCA1 in acute myeloid leukemia. *Oncotarget* **6,** 18534–18544 (2015).
295. Jolma, A. *et al.* DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527,** 384–388 (2015).
296. Kume, T. *et al.* The forkhead/winged helix gene Mf1 is disrupted in the pleiotropic mouse mutation congenital hydrocephalus. *Cell* **93,** 985–996 (1998).
297. Tümer, Z. & Bach-Holm, D. Axenfeld-Rieger syndrome and spectrum of PITX2 and FOXC1 mutations. *Eur. J. Hum. Genet. EJHG* **17,** 1527–1539 (2009).
298. Nishimura, D. Y. *et al.* The forkhead transcription factor gene FKHL7 is responsible for glaucoma phenotypes which map to 6p25. *Nat. Genet.* **19,** 140–147 (1998).
299. Lay, K., Kume, T. & Fuchs, E. FOXC1 maintains the hair follicle stem cell niche and governs stem cell quiescence to preserve long-term tissue-regenerating potential. *Proc. Natl. Acad. Sci.* **113,** E1506–E1515 (2016).
300. Somerville, T. D. D. *et al.* Frequent Derepression of the Mesenchymal Transcription Factor Gene FOXC1 in Acute Myeloid Leukemia. *Cancer Cell* **28,** 329–342 (2015).
301. Omatsu, Y., Seike, M., Sugiyama, T., Kume, T. & Nagasawa, T. Foxc1 is a critical regulator of haematopoietic stem/progenitor cell niche formation. *Nature* **508,** 536–540 (2014).
302. Wang, J. *et al.* FOXC1 promotes melanoma by activating MST1R/PI3K/AKT. *Oncotarget* **7,** 84375–84387 (2016).
303. Xia, L. *et al.* Overexpression of forkhead box C1 promotes tumor metastasis and indicates poor prognosis in hepatocellular carcinoma. *Hepatol. Baltim. Md* **57,** 610–624 (2013).
304. Wang, L. *et al.* High level of FOXC1 expression is associated with poor prognosis in pancreatic ductal adenocarcinoma. *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.* **34,** 853–858 (2013).
305. Ray, P. S. *et al.* Basal-like breast cancer defined by FOXC1 expression offers superior prognostic value: a retrospective immunohistochemical study. *Ann. Surg. Oncol.* **18,** 3839–3847 (2011).
306. Han, B. *et al.* FOXC1: an emerging marker and therapeutic target for cancer. *Oncogene* **36,** 3957–3963 (2017).
307. Thorsteinsdottir, U. *et al.* Overexpression of the myeloid leukemia-associated Hoxa9 gene in bone marrow cells induces stem cell expansion. *Blood* **99,** 121–129 (2002).
308. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma. Oxf. Engl.* **30,** 2114–2120 (2014).
309. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11,** 1650–1667 (2016).

310. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47 (2015).
311. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2008).
312. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
313. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
314. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38,** 576–589 (2010).
315. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
316. Saldanha, A. J. Java Treeview--extensible visualization of microarray data. *Bioinforma. Oxf. Engl.* **20,** 3246–3248 (2004).
317. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27,** 3423–3424 (2011).
318. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14,** 178–192 (2013).
319. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).
320. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
321. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25,** 2078–2079 (2009).
322. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).
323. Zhu, L. J. *et al.* ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* **11,** 237 (2010).
324. Chen, H. & Boutros, P. C. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* **12,** 35 (2011).
325. Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72,** 248–254 (1976).
326. Bert, A. G., Burrows, J., Osborne, C. S. & Cockerill, P. N. Generation of an improved luciferase reporter gene plasmid that employs a novel mechanism for high-copy replication. *Plasmid* **44,** 173–182 (2000).
327. Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* **32,** 279–284 (2014).
328. Engler, C., Kandzia, R. & Marillonnet, S. A One Pot, One Step, Precision Cloning Method with High Throughput Capability. *PLoS ONE* **3**, e3647 (2008).
329. Argiropoulos, B. & Humphries, R. K. Hox genes in hematopoiesis and leukemogenesis. *Oncogene* **26,** 6766–6776 (2007).
330. Kroon, E. *et al.* Hoxa9 transforms primary bone marrow cells through specific collaboration with Meis1a but not Pbx1b. *EMBO J.* **17,** 3714–3725 (1998).

331. Shanmugam, K., Green, N. C., Rambaldi, I., Saragovi, H. U. & Featherstone, M. S. PBX and MEIS as Non-DNA-Binding Partners in Trimeric Complexes with HOX Proteins. *Mol. Cell. Biol.* **19,** 7577–7588 (1999).
332. Wong, P., Iwasaki, M., Somervaille, T. C. P., So, C. W. E. & Cleary, M. L. Meis1 is an essential and rate-limiting regulator of MLL leukemia stem cell potential. *Genes Dev.* **21,** 2762–2774 (2007).
333. Argiropoulos, B., Yung, E. & Humphries, R. K. Unraveling the crucial roles of Meis1 in leukemogenesis and normal hematopoiesis. *Genes Dev.* **21,** 2845–2849 (2007).
334. Zhu, J. & Paul, W. E. CD4 T cells: fates, functions, and faults. *Blood* **112,** 1557–1569 (2008).
335. FANTOM Consortium and the RIKEN PMI and CLST (DGT) *et al.* A promoter-level mammalian expression atlas. *Nature* **507,** 462–470 (2014).
336. Pang, S. H. M. *et al.* PU.1 cooperates with IRF4 and IRF8 to suppress pre-B-cell leukemia. *Leukemia* **30,** 1375–1387 (2016).
337. Liebermann, D. A., Gregory, B. & Hoffman, B. AP-1 (Fos/Jun) transcription factors in hematopoietic differentiation and apoptosis. *Int. J. Oncol.* **12,** 685–1385 (1998).
338. Piper, J. *et al.* Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.* **41,** e201 (2013).
339. Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36,** D480–D484 (2008).
340. Yanai, H., Negishi, H. & Taniguchi, T. The IRF family of transcription factors. *Oncoimmunology* **1,** 1376–1386 (2012).
341. Taniguchi, T., Ogasawara, K., Takaoka, A. & Tanaka, N. IRF family of transcription factors as regulators of host defense. *Annu. Rev. Immunol.* **19,** 623–655 (2001).
342. Yusuf, I. *et al.* KLF4 is a FOXO target gene that suppresses B cell proliferation. *Int. Immunol.* **20,** 671–681 (2008).
343. Greer, E. L. & Brunet, A. FOXO transcription factors at the interface between longevity and tumor suppression. *Oncogene* **24,** 7410–7425 (2005).
344. Trojer, P. *et al.* L3MBTL2 protein acts in concert with PcG protein-mediated monoubiquitination of H2A to establish a repressive chromatin structure. *Mol. Cell* **42,** 438–450 (2011).
345. Gopisetty, G., Xu, J., Sampath, D., Colman, H. & Puduvalli, V. K. Epigenetic regulation of CD133/PROM1 expression in glioma stem cells by Sp1/myc and promoter methylation. *Oncogene* **32,** 3119–3129 (2013).
346. Katoh, Y. & Katoh, M. Comparative genomics on PROM1 gene encoding stem cell marker CD133. *Int. J. Mol. Med.* **19,** 967–970 (2007).
347. Boss, C. N. *et al.* Identification and characterization of T-cell epitopes deduced from RGS5, a novel broadly expressed tumor antigen. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **13,** 3347–3355 (2007).
348. Suknuntha, K. *et al.* Discovery of survival factor for primitive chronic myeloid leukemia cells using induced pluripotent stem cells. *Stem Cell Res.* **15,** 678–693 (2015).
349. Obulkasim, A., Fornerod, M., Zwaan, M. C., Reinhardt, D. & van den Heuvel-Eibrink, M. M. Subtype prediction in pediatric acute myeloid leukemia: classification using differential network rank conservation revisited. *BMC Bioinformatics* **16,** (2015).

350. Qin, J. *et al.* The Polycomb Group Protein L3mbtl2 Assembles an Atypical PRC1-Family Complex that Is Essential in Pluripotent Stem Cells and Early Development. *Cell Stem Cell* **11,** 319–332 (2012).
351. Amin, S. *et al.* Hoxa2 selectively enhances Meis binding to change a branchial arch ground state. *Dev. Cell* **32,** 265–277 (2015).
352. Bowers, S. R. *et al.* A Conserved Insulator That Recruits CTCF and Cohesin Exists between the Closely Related but Divergently Regulated Interleukin-3 and Granulocyte-Macrophage Colony-Stimulating Factor Genes. *Mol. Cell. Biol.* **29,** 1682–1693 (2009).
353. Kaucká, M. *et al.* The planar cell polarity pathway drives pathogenesis of chronic lymphocytic leukemia by the regulation of B-lymphocyte migration. *Cancer Res.* **73,** 1491–1501 (2013).
354. Secchiero, P. *et al.* The expression levels of the pro-apoptotic XAF-1 gene modulate the cytotoxic response to Nutlin-3 in B chronic lymphocytic leukemia. *Leukemia* **24,** 480–483 (2009).
355. Edwards, S. V. & Hedrick, P. W. Evolution and ecology of MHC molecules: from genomics to sexual selection. *Trends Ecol. Evol.* **13,** 305–311 (1998).
356. Alharbi, R. A., Pettengell, R., Pandha, H. S. & Morgan, R. The role of HOX genes in normal hematopoiesis and acute leukemia. *Leukemia* **27,** 1000–1008 (2013).
357. Mebius, R. E. & Kraal, G. Structure and function of the spleen. *Nat. Rev. Immunol.* **5,** 606–616 (2005).
358. Somervaille, T. C. P. & Cleary, M. L. Identification and characterization of leukemia stem cells in murine MLL-AF9 acute myeloid leukemia. *Cancer Cell* **10,** 257–268 (2006).
359. Wells, S. M., Kantor, A. B. & Stall, A. M. CD43 (S7) expression identifies peripheral B cell subsets. *J. Immunol.* **153,** 5503–5515 (1994).
360. Simmons, D. L., Tan, S., Tenen, D. G., Nicholson-Weller, A. & Seed, B. Monocyte antigen CD14 is a phospholipid anchored membrane protein. *Blood* **73,** 284–289 (1989).
361. Penna, G. *et al.* Cutting edge: differential chemokine production by myeloid and plasmacytoid dendritic cells. *J. Immunol. Baltim. Md 1950* **169,** 6673–6676 (2002).
362. Takahashi, H. *et al.* An essential role of macrophage inflammatory protein 1α/CCL3 on the expression of host's innate immunities against infectious complications. *J. Leukoc. Biol.* **72,** 1190–1197 (2002).
363. Schoenemeyer, A. *et al.* The interferon regulatory factor, IRF5, is a central mediator of toll-like receptor 7 signaling. *J. Biol. Chem.* **280,** 17005–17012 (2005).
364. Hayden, M. S., West, A. P. & Ghosh, S. NF-kappaB and the immune response. *Oncogene* **25,** 6758–6780 (2006).
365. Parameswaran, N. & Patial, S. Tumor Necrosis Factor-α Signaling in Macrophages. *Crit. Rev. Eukaryot. Gene Expr.* **20,** 87–103 (2010).
366. Yamanaka, R. *et al.* Impaired granulopoiesis, myelodysplasia, and early lethality in CCAAT/enhancer binding protein epsilon-deficient mice. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 13187–13192 (1997).
367. Akasaka, T. *et al.* Five members of the CEBP transcription factor family are targeted by recurrent IGH translocations in B-cell precursor acute lymphoblastic leukemia (BCP-ALL). *Blood* **109,** 3451–3461 (2007).
368. Gardiner, J. D. *et al.* C/EBPβ-1 promotes transformation and chemoresistance in Ewing sarcoma cells. *Oncotarget* **8,** 26013–26026 (2017).
369. Alberich-Jordà, M. *et al.* C/EBPγ deregulation results in differentiation arrest in acute myeloid leukemia. *J. Clin. Invest.* **122,** 4490–4504 (2012).

370. Shackleford, T. J. *et al.* Stat3 and CCAAT/enhancer binding protein beta (C/EBP-beta) regulate Jab1/CSN5 expression in mammary carcinoma cells. *Breast Cancer Res.* **13,** R65 (2011).
371. Lieu, Y. K. & Reddy, E. P. Conditional c-myb knockout in adult hematopoietic stem cells leads to loss of self-renewal due to impaired proliferation and accelerated differentiation. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 21689–21694 (2009).
372. Ramsay, R. G. & Gonda, T. J. MYB function in normal and cancer cells. *Nat. Rev. Cancer* **8,** 523–534 (2008).
373. Lahortiga, I. *et al.* Duplication of the MYB oncogene in T cell acute lymphoblastic leukemia. *Nat. Genet.* **39,** 593–595 (2007).
374. Hess, J. L. *et al.* c-Myb is an essential downstream target for homeobox-mediated transformation of hematopoietic cells. *Blood* **108,** 297–304 (2006).
375. Shen, W. F. *et al.* AbdB-like Hox proteins stabilize DNA binding by the Meis1 homeodomain proteins. *Mol. Cell. Biol.* **17,** 6448–6458 (1997).
376. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci.* **107,** 21931–21936 (2010).
377. Ross-Innes, C. S., Brown, G. D. & Carroll, J. S. A co-ordinated interaction between CTCF and ER in breast cancer cells. *BMC Genomics* **12,** 593 (2011).
378. Lee, J., Krivega, I., Dale, R. K. & Dean, A. The LDB1 Complex Co-opts CTCF for Erythroid Lineage-Specific Long-Range Enhancer Interactions. *Cell Rep.* **19,** 2490–2502 (2017).
379. Iwafuchi-Doi, M. *et al.* The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. *Mol. Cell* **62,** 79–91 (2016).
380. Vegi, N. M. *et al.* MEIS2 Is an Oncogenic Partner in AML1-ETO-Positive AML. *Cell Rep.* **16,** 498–507 (2016).
381. Berry, F. B., Saleem, R. A. & Walter, M. A. FOXC1 Transcriptional Regulation Is Mediated by N- and C-terminal Activation Domains and Contains a Phosphorylated Transcriptional Inhibitory Domain. *J. Biol. Chem.* **277,** 10292–10297 (2002).
382. Saleem, R. A., Banerjee-Basu, S., Berry, F. B., Baxevanis, A. D. & Walter, M. A. Analyses of the effects that disease-causing missense mutations have on the structure and function of the winged-helix protein FOXC1. *Am. J. Hum. Genet.* **68,** 627–641 (2001).
383. Shi, J. *et al.* The Polycomb complex PRC2 supports aberrant self-renewal in a mouse model of MLL-AF9;NrasG12D acute myeloid leukemia. *Oncogene* **32,** 930–938 (2013).
384. Kume, T., Jiang, H., Topczewska, J. M. & Hogan, B. L. M. The murine winged helix transcription factors, Foxc1 and Foxc2, are both required for cardiovascular development and somitogenesis. *Genes Dev.* **15,** 2470–2482 (2001).
385. Fulda, S. Inhibitor of apoptosis proteins in hematological malignancies. *Leukemia* **23,** 467–476 (2009).
386. Elias, S. *et al.* Immune evasion by oncogenic proteins of acute myeloid leukemia. *Blood* **123,** 1535–1543 (2014).
387. Heuser, M. *et al.* MN1 overexpression induces acute myeloid leukemia in mice and predicts ATRA resistance in patients with AML. *Blood* **110,** 1639–1647 (2007).
388. Heuser, M. *et al.* Cell of origin in AML: Susceptibility to MN1-induced transformation is regulated by the MEIS1/AbdB-like HOX protein complex. *Cancer Cell* **20,** 39–52 (2011).

389. Meyer, S. E. *et al.* DNMT3A Haploinsufficiency Transforms FLT3ITD Myeloproliferative Disease into a Rapid, Spontaneous, and Fully Penetrant Acute Myeloid Leukemia. *Cancer Discov.* **6,** 501–515 (2016).
390. Stevens, M. M. *et al.* Drug sensitivity of single cancer cells is predicted by changes in mass accumulation rate. *Nat. Biotechnol.* **34,** 1161–1167 (2016).
391. Hirose, M. *et al.* A novel monocytoid cultured cell line, P31/Fujioka, derived from acute monoblastic leukemia. *Gan* **73,** 735–741 (1982).
392. Rao, V. N., Ohno, T., Prasad, D. D., Bhattacharya, G. & Reddy, E. S. Analysis of the DNA-binding and transcriptional activation functions of human Fli-1 protein. *Oncogene* **8,** 2167–2173 (1993).
393. Li, Y., Luo, H., Liu, T., Zacksenhaus, E. & Ben-David, Y. The ets transcription factor Fli-1 in development, cancer and disease. *Oncogene* **34,** 2022–2031 (2015).
394. Lozzio, B. *et al.* A Multipotential Leukemia Cell Line (K-562) of Human Origin. *Exp. Biol. and Med.* **166** (1981).
395. Andersson, L. C., Nilsson, K. & Gahmberg, C. G. K562—A human erythroleukemic cell line. *Int. J. Cancer* **23,** 143–147 (1979).
396. Armstrong, J. A. & Emerson, B. M. NF-E2 disrupts chromatin structure at human beta-globin locus control region hypersensitive site 2 in vitro. *Mol. Cell. Biol.* **16,** 5634–5644 (1996).
397. Lichtinger, M. *et al.* RUNX1 reshapes the epigenetic landscape at the onset of haematopoiesis. *EMBO J.* **31,** 4318–4333 (2012).
398. Orkin, S. H. & Zon, L. I. Hematopoiesis: An Evolving Paradigm for Stem Cell Biology. *Cell* **132,** 631–644 (2008).
399. Rudra, D. *et al.* Transcription factor Foxp3 and its protein partners form a complex regulatory network. *Nat. Immunol.* **13,** 1010–1019 (2012).
400. van der Vos, K. E. & Coffer, P. J. FOXO-binding partners: it takes two to tango. *Oncogene* **27,** 2289–2299 (2008).
401. Hurtado, A., Holmes, K. A., Ross-Innes, C. S., Schmidt, D. & Carroll, J. S. FOXA1 is a critical determinant of Estrogen Receptor function and endocrine response. *Nat. Genet.* **43,** 27–33 (2011).
402. Pomerantz, M. M. *et al.* The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. *Nat. Genet.* **47,** 1346–1351 (2015).
403. Zhao, Y., Tindall, D. J. & Huang, H. Modulation of Androgen Receptor by FOXA1 and FOXO1 Factors in Prostate Cancer. *Int. J. Biol. Sci.* **10,** 614–619 (2014).
404. Mieczkowski, J. *et al.* MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. *Nat. Commun.* **7,** ncomms11485 (2016).
405. Volpe, G. *et al.* Prognostic significance of high GFI1 expression in AML of normal karyotype and its association with a FLT3-ITD signature. *Sci. Rep.* **7,** 11148 (2017).
406. Tzelepis, K. *et al.* A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. *Cell Rep.* **17,** 1193–1205 (2016).
407. Network, T. C. G. A. R. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *N. Engl. J. Med.* **368,** 2059–2074 (2013).
408. Kubonishi, I. *et al.* Two new human myeloid cell lines derived from acute promyelocytic leukemia and chronic myelocytic leukemia. *Gan* **74,** 319–322 (1983).

409. Xia, N. *et al.* Role of SIRT1 and FOXO factors in eNOS transcriptional activation by resveratrol. *Nitric Oxide* **32,** 29–35 (2013).
410. Meredith, D. M. *et al.* Program specificity for Ptf1a in pancreas versus neural tube development correlates with distinct collaborating cofactors and chromatin accessibility. *Mol. Cell. Biol.* **33,** 3166–3179 (2013).
411. Tsuchiya, S. *et al.* Induction of maturation in cultured human monocytic leukemia cells by a phorbol diester. *Cancer Res.* **42,** 1530–1536 (1982).
412. Wilson, L., Butcher, C. J. & Kellie, S. Calcium ionophore A23187 induces interleukin-8 gene expression and protein secretion in human monocytic cells. *FEBS Lett.* **325,** 295–298 (1993).
413. Zenke, M. *et al.* Multiple sequence motifs are involved in SV40 enhancer function. *EMBO J.* **5,** 387–397 (1986).
414. Kelly, J. J. & Wildeman, A. G. Role of the SV40 enhancer in the early to late shift in viral transcription. *Nucleic Acids Res.* **19,** 6799–6804 (1991).
415. Goldfarb, A. N., Lewandowska, K. & Shoham, M. Determinants of helix-loop-helix dimerization affinity. Random mutational analysis of SCL/tal. *J. Biol. Chem.* **271,** 2683–2688 (1996).
416. Cereghini, S. & Yaniv, M. Assembly of transfected DNA into chromatin: structural changes in the origin-promoter-enhancer region upon replication. *EMBO J.* **3,** 1243–1253 (1984).
417. Bishop, C., Ramalho, M., McBride, A. A., Higgins, C. F. & Krauzewicz, N. 1150. Understanding Transcriptional Regulation of Exogenously Delivered DNA. *Mol. Ther.* **7,** S443 (2003).
418. Dickel, D. E. *et al.* Function-based identification of mammalian enhancers using site-specific integration. *Nat. Methods* **11,** 566–571 (2014).
419. Crawford, G. E. *et al.* Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 992–997 (2004).
420. Swinstead, E. E. *et al.* Steroid Receptors Reprogram FoxA1 Occupancy through Dynamic Chromatin Transitions. *Cell* **165,** 593–605 (2016).
421. Nagarajan, S. *et al.* Bromodomain Protein BRD4 Is Required for Estrogen Receptor-Dependent Enhancer Activation and Gene Transcription. *Cell Rep.* **8,** 460–469 (2014).
422. Shi, J. *et al.* Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation. *Genes Dev.* (2013). doi:10.1101/gad.232710.113
423. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* **6,** e21856 (2017).
424. Liu, X. *et al.* In Situ Capture of Chromatin Interactions by Biotinylated dCas9. *Cell* **170,** 1028–1043.e19 (2017).
425. Fu, Z. & Tindall, D. J. FOXOs, cancer and regulation of apoptosis. *Oncogene* **27,** 2312–2319 (2008).
426. Sadasivam, S., Duan, S. & DeCaprio, J. A. The MuvB complex sequentially recruits B-Myb and FoxM1 to promote mitotic gene expression. *Genes Dev.* **26,** 474–489 (2012).
427. Wang, J. *et al.* FOXC1 regulates the functions of human basal-like breast cancer cells by activating NF-κB signaling. *Oncogene* **31,** 4798–4802 (2012).
428. Xu, Y. *et al.* Overexpression of FOXC1 correlates with poor prognosis in gastric cancer patients. *Histopathology* **64,** 963–970 (2014).

429. Xu, Y.-Y., Tian, J., Hao, Q. & Yin, L.-R. MicroRNA-495 downregulates FOXC1 expression to suppress cell growth and migration in endometrial cancer. *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.* **37,** 239–251 (2016).
430. Illendula, A. *et al.* Small Molecule Inhibitor of CBFβ-RUNX Binding for RUNX Transcription Factor Driven Cancers. *EBioMedicine* **8,** 117–131 (2016).
431. Somerville, T. D. D. & Somervaille, T. C. P. Tissue-inappropriate derepression of FOXC1 is frequent and functional in human acute myeloid leukemia. *Mol. Cell. Oncol.* **3,** (2016).