THE UNIVERSITY OF CHICAGO

THE EPIGENETIC REGULATION MECHANISM OF CUX1 IN HEMATOPOIESIS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

COMMITTEE ON CANCER BIOLOGY

BY

WEIHAN LIU

CHICAGO, ILLINOIS

JUNE 2024

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I wish to express my profound gratitude to my advisor, Dr. Megan McNerney. Venturing into graduate school with a keen interest in applying computational biology to cancer research and possessing modest programming skills, the feasibility of my dream was uncertain. Dr. McNerney's steadfast support and confidence have been instrumental in my pursuit of a computational project, for which I am eternally grateful. Her consistent encouragement to explore new ideas, coupled with candid feedback, has been pivotal in keeping my endeavors on course. Her support extended to facilitating my applications for multiple grants and scholarships, as well as providing opportunities to mentor undergraduate students. In my third year, when my initial single-cell omics project encountered a bottleneck, Dr. McNerney's guidance was invaluable in transitioning to another project. A significant takeaway from my graduate school experience is Dr. McNerney's leadership ethos—her approachability, inclusivity, and respect for all team members, irrespective of their rank.

A special thank-you is also due to Dr. Angela Stoddart for her unwavering mentorship and collaboration. Dr. Stoddart's readiness to assist, from providing thesis feedback to educating me on new biological concepts, has been a cornerstone of my academic progress.

My committee members, Dr. Xin He, Dr. Barbara Kee, and Dr. Alexander Pearson, deserve a heartfelt thank you. Your guidance and suggestions have been integral to my thesis and have helped me to navigate the complex paths of research. Your willingness to offer guidance beyond the committee meetings and to connect me with domain experts has been truly invaluable.

I am grateful for the collective spirit and support of the McNerney lab members. Molly Imgruet, my mentor during my rotation, was always available for my queries. Jeremy

Baeten collaborated with me on the TSG machine learning project, bringing joy and knowledge to our work. Jeff Kurkewich, my office mate and collaborator for the CUX1 pioneer factor project, brightened every day with his presence and shared his expertise in epigenomics. Stephanie Konecki collaborated with me on the single-cell RNA-sequencing project and was instrumental in my development as a computational biologist. Saira Khan, the wet lab expert, provided invaluable experimental assistance. Ningfei An, Matt Jotte, Tanner Martinez, Joseph Cannova, Madhavi Senagolage, Raven Moten, Katarzyna Zawieracz, Yuqing Xue, Lia Jueng, Dhivyaa Anandan, Henna Nam, and others have made my PhD journey a fulfilling experience. I also extend my thanks to Dr. Steve Kron and Don Wolfgeher for their support in proteomics analysis.

Outside the lab, my friends Manny Rocha, Chang Cui, Wenchao Liu, Anqi Yu, Avelino De Leon, Julian Lutze, Peter Mintun, Justin Chang, Manas Rajaram, Alberto Tohme, Aileen Brown-Cuevas, Andrew Leonard, and Daniel Guoshuai Cao have made life enjoyable both in and out of school. A special shout-out to Shu Fu, my long-time friend and neighbor, for his steadfast companionship.

Lastly, my family has been my greatest support. The unconditional love and backing from my parents, the discipline and attitude instilled by my grandfather, and the frugality and pragmatism taught by my grandmother have been my guiding lights. My fiancée and soon-to-be wife Audrey offer me unparallel support and I am extremely lucky to have you in my life. And to my cat H3K27ac, thank you for your fluffy companionship and enduring the uniqueness of your name.

This acknowledgment reflects my deepest appreciation for everyone who has supported and inspired me throughout my academic journey. Thank you all.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CIN - Chromosomal instability

AML - Acute myeloid leukemia

CML - Chronic myeloid leukemia

ML – Machine learning

TF - Transcription factor

PF - Pioneer factor

HSPC: Hematopoietic stem and progenitor cell

HSC: Hematopoietic stem cell

ChIP-seq - Chromatin immunoprecipitation sequencing

ATAC-seq - Assay for transposase-accessible chromatin with sequencing

RNA-seq – RNA sequencing

scRNA-seq – single cell RNA sequencing

GSEA – Gene set enrichment analysis

GO – Gene ontology

co-IP mass spec: co-immunoprecipitation mass spectrometry

AAV: Adeno-associated virus

AAVS1: Adeno-associated virus integration site 1

DDR: DNA damage repair

HAT: histone acetyltransferases

TSS: Transcription start site

ABSTRACT

This three-parts thesis presents discovery of the molecular mechanisms underpinning hematopoiesis and its dysregulation in cancer. Identifying tumor suppressor genes on chromosome regions affected by aneuploidy has been historically challenging due to the large number of genes involved. The first part of the thesis leveraged published genome-wide perturbation screen data and advancement in machine learning algorithms in recent years. This work led to a supervised machine learning workflow that systemically predicted the tumor suppressor gene-like activities for all chromosome 7 genes. The second and third parts focus on the multifaceted roles of CUX1 as a pioneer transcription factor. CUX1 is a homeodomain-containing transcription factor (TF) that is essential for development and differentiation of multiple tissues. CUX1 is recurrently mutated or deleted in cancer, particularly in myeloid malignancies. However, the mechanisms by which CUX1 regulates gene expression and differentiation remain poorly understood, creating a barrier to understanding the tumor suppressive functions of CUX1. Herein, we demonstrate that CUX1 directs the BAF chromatin remodeling complex to DNA to increase DNA accessibility in hematopoietic cells. CUX1 preferentially regulates lineage-specific enhancers, and CUX1 target genes are predictive of cell fate *in vivo*. Moreover, the thesis illuminates the intricate relationship between CUX1 and GATA1, two key regulators in erythropoiesis. In erythroid differentiation, CUX1 dynamically shifts binding targets from hematopoietic stem cell (HSC) -specific enhancers to erythroid specific enhancers co-bound by GATA1. CUX1 gatekeeps GATA1 from abnormal and promiscuous binding by direct physical shielding and indirect mechanisms. These data indicate that CUX1 possesses pioneer factor activities to epigenetically regulate hematopoietic lineage commitment and homeostasis. CUX1 deficiency disrupts these processes in stem and progenitor cells, facilitating transformation. In the erythroid branch of hematopoiesis, CUX1 promotes healthy differentiation through

ensuring proper GATA1 binding. By bridging molecular insights from aneuploidy, CUX1 epigenetic regulatory mechanisms, and CUX1-GATA1 interaction, this thesis provides novel insights of the molecular machinery governing hematopoiesis and offers novel perspectives on cancer biology and treatment strategies.

CHAPTER 1: INTRODUCTION

1.1 Aneuploidy in cancer

Chromosomal instability (CIN), manifested as structural or numerical chromosomal abnormalities, is a fundamental hallmark of cancer. Structural chromosome rearrangements have been extensively investigated and shown to play important roles in tumorigenesis by activating oncogenes and deactivating tumor suppressor genes (TSG).[1] A classic example is the chromosome translocation that creates the *BCR-ABL* oncogene that drives chronic myeloid leukemia (CML).[2] The discovery subsequently led to the development of the life-saving tyrosine kinase inhibitor, imatinib. Despite progress in understanding structural chromosome rearrangement, the significance of numerical chromosome alterations, known as aneuploidy, in tumor development is less well understood.

Aneuploidy, induced by CIN, was initially defined as numerical aberration across the entire chromosome.[3] However, recent pan-cancer analyses have expanded this definition to include gains or losses of chromosome arms, a condition termed "partial aneuploidy".[4] The phenomenon of aneuploidy was first observed by David von Hansemann in the late 19th century in epithelial tumor cells and was further studied by Theodore Boveri in the early 20th century through experiments in sea urchin eggs, leading to the hypothesis that malignant tumors may arise from abnormal chromosomal constitutions.[5] Aneuploidy primarily arises from errors in chromosome segregation processes, with merotelic attachments, spindle assembly checkpoint failures, and chromosome cohesion defects as principal mechanisms.[6,7]

Aneuploidy is frequently detected in cancer, with a widespread presence in various blood and solid tumor types.[8–11] The Mitelman Database reports that nearly 90% of all solid tumors and 50% of blood cancers display aneuploidy. Large-scale DNA copy number

analyses show that up to a quarter of the typical cancer cell genome is affected by whole-chromosome or whole arm somatic copy number alterations.[12,13]

For healthy cells, aneuploidy generally has a detrimental effect on the fitness by damaging essential pathways such as proliferation, DNA damage repair and cellular metabolism.[13–15] However, for cancer cells, the effect of aneuploidy on cancer is context dependent. Aneuploidy could either promote or suppress tumorigenesis depending on tumor type, stage, type of genes on the affected regions, immune interactions, and tumor microenvironment.[16]

On the one hand, studies have shown that introducing extra copies of chromosomes could function as tumor suppressors,[17] and a pan-cancer genome-wide analysis showed that frequency of aneuploidy is inversely correlated with number of coding genes on the chromosomal regions affected,[12] suggesting that aneuploidy could carry a fitness penalty for cancer cells.

On the other hand, aneuploidy usually promotes genome instability, which is the substrate for malignant transformation.[18] There is abundant evidence suggesting aneuploidy promotes tumorigenesis. For example, aneuploidy could confer a survival advantage to cancer cells under conditions of stress *in vitro*,[19] drives tumorigenesis by inactivating TSGs or overexpressing oncogenes,[20,21] and promotes epithelial-to-mesenchymal transition and metastasis by activating the cGAS-cGAMP-STING pathway.[22] Furthermore, emerging evidence showed that aneuploidy might be exploited by cancer cells to delay cell cycle and negatively impact drug metabolism processes.[23–25] Clinically, aneuploidy has been used as a prognostic marker for various cancer types. High levels of aneuploidy are generally

associated with poorer prognosis in most human cancer, while only a small proportion of aneuploidy is associated with better prognosis.[16,26,27]

Despite these widespread implications of aneuploidy in cancer, a formidable challenge is to understand the molecular mechanisms of individual genes on the chromosome regions impacted by aneuploidy. This has been difficult for primarily two reasons. First, the impacted chromosome or chromosome arm regions typically contain hundreds or even thousands of genes, and the genes on these regions could act cooperatively. This complexity makes elucidation of individual gene's function difficult. Second, under certain contexts, it is unclear whether aneuploidy drives oncogenesis, or is merely a passenger effect due to the severe dysregulation of cancer cells.[10] Despite all these challenges, elucidating the roles of specific genes and pathways affected by aneuploidy in specific cancers is crucial to understand how cancers form and generate new therapeutic strategies. Advancement in precise genetic manipulation tools such as RNAi and CRISPR, and high throughput genetic screens have vastly enabled our ability to interrogate the phenotypical contribution of specific aneuploidy-associated genes, through either gain-of-function or loss-of-function screens. For the first time, we can identify putative oncogenes or tumor suppressors by deleting or knocking in a copy of the gene and investigate the effect on cellular phenotypes such as proliferation, for hundreds and even thousands of genes at the same time.

1.2 -7/del(7q) in myeloid malignancies

Aneuploidy-associated chromosomal duplications, deletions and translocations have frequently been reported in hematopoietic malignancies. In fact, the first chromosomal translocation associated with the pathogenesis of cancer was identified in hematopoietic cells.[28] The loss of all chromosome 7 (Monosomy 7 or -7) or the long arm (del(7q)) is one of

the most recognized high risk cytogenetic abnormalities in pediatric and adult myeloid malignancies.[29] (**Figure 1.1, Table 1**).

Table 1: The prevalence of -7/del7(q) in myeloid malignancies

Cancer type	**% cases with -7/del(7q)**	**Source**
De novo acute myeloid leukemia (AML)	10-15%	PMID: 27276561
Myelodysplastic Syndrome (MDS)-Adults	14%	PMID: 18414863
Myelodysplastic Syndrome (MDS)-Pediatric	40%	PMID: 29146900
Chronic Myelomonocytic Leukemia (CMML)	14%	PMID: 11806985
Juvenile Myelomonocytic Leukemia (JMML)	33%	PMID: 10086728
Myeloid Neoplasms-post cytotoxic therapy (MN-pCT)	50%	PMID: 12623843

Figure 1.1. Karyotype of normal chromosome 7 (top), monosomy 7 (or -7, bottom left), and del7(q) (Bottom right)

-7/del(7q) was first observed as a frequent cytogenetic abnormality in AML more than half a century ago.[30] -7/del(7q) is detected in clonal hematopoiesis and found to be initiating events in malignant transformation.[31,32] -7/del(7q) has been shown to be involved in diseases including bone marrow failure, myelodysplastic syndrome (MDS), aplastic anemia, acute myeloid leukemia (AML) and therapy related myeloid neoplasms (t-MN).[33–37] Clinically, -7/del(7q) is considered an adverse prognostic event associated with higher risk MDS, faster transformation to AML and poor overall survival in AML.[38,39] There is also evidence suggesting -7/del(7q) as the driver event in myeloid diseases. For example, a study found ~30% of pediatric MDS patients with -7/del(7q) had no other detected cytogenetic abnormalities in the genomic coding region, indicating -7/del(7q) alone might be an early driver in disease development.[40]

Despite the important clinical implications of -7/del(7q), the underlying mechanism that promotes transformation remained elusive. Key TSGs were postulated to reside on chromosome 7. However, identifying these genes has been difficult. Besides the complexity associated with the sheer number of genes and the intertwined interactions among them, how TSGs function also adds another layer of complexity. Conventional thinking dictates that both alleles of a TSG need to be inactivated through mutation or epigenetic silencing, in order to cause a phenotypical consequence (Knudsen's two-hit hypothesis).[41] Studies in recent decades also showed that many TSGs function in a haploinsufficient manner, meaning loss of just one copy of a gene, through either mutation or aneuploidy, is sufficient to drive diseases.[42,43] Notably, a "second-hit" mutation on chromosome 7 genes is not recurrently observed for genes impacted by -7/del(7q), except *EZH2*,[44] suggesting that most of these TSGs encoded on chromosome 7 are haploinsufficient. An additional layer of complexity is that loss of multiple adjacent genes could result in a combinatorial effect, termed contiguous

gene syndrome.[45,46] The combinatorial decrease of gene expression is thought to be more pathogenic than the loss of a single gene within the segment. Contiguous gene syndrome has been observed in myeloid malignancies. For example, in another common cytogenetic abnormality event del(5q), haploinsufficiency of both *Apc* and *Egr1* is necessary to cause myeloid malignancy in *Trp53* low mice.[47] Another study found that co-suppression of multiple genes on chromosome 8p synergistically promote tumor growth in mice and is associated with worse survival than loss of any individual genes.[48]

There have been extensive efforts to identify TSGs on chromosome 7 and elucidate the pathogenic mechanisms of these TSGs in myeloid malignancies. For example, *CUX1*, a homeobox-containing transcription factor encoded on 7q22, was identified through mapping commonly deleted regions, and shown to function as a TSG.[49] *CUX1* will be explored more in detail in the next section. 7q35-36 and 7q34 are two other commonly deleted regions reported on chromosome 7.[38,50] Several candidate tumor suppressor genes have been identified on these regions. For example. *EZH2*, which encodes for a H3K27 methyltransferase, was identified on 7q36.1 by searching for second-hit mutations. *EZH2* is a component of the Polycomb Repressive Complex 2 (PRC2) that epigenetically represses genes involved in stem cell fate determination and is frequently mutated in myeloid malignancies.[44] *MLL3*, a H3K4 methyltransferase on 7q36.1 was identified as a haploinsufficient TSG that cooperates with reduced expression of *Nf1* and *Trp53* to promote leukemogenesis in mice.[51] Other putative chromosome 7 TSGs identified so far include *LUC7L2*, which is a pre-mRNA splicing factor component encoded on 7q34.[52] *CUL1 (*on 7q36.1*)*, which is a E3 ubiquitin ligase complex component, is frequently mutated in myeloid neoplasms.[52,53] Homozygous or heterozygous deletion of *SAMD9L* on 7q21.2 was shown to cause mice to develop bone marrow failure and dysplasia akin to human with diseases with -

7/del(7q).[54] However, besides these few candidate TSGs, there is no systemic study to identify the TSG activities on chromosome 7. Uncovering the latent TSG activities of chromosome 7 genes could provide novel therapeutic targets for myeloid malignancies.

1.3 Machine learning in cancer research

Advancement in next generation sequencing technologies ushered biology into the big data era. High dimensional "omics" data including DNA, RNA sequencing, proteomics, metabolomics and epigenomics data often involved tens of thousands and even millions of entries (e.g. single cells, genes in RNA-seq, genomic loci etc) across multiple comparison groups and various perturbation conditions. Very often, subtle patterns are buried underneath haystacks of data and are impossible to detect by human eyes or using simple statistical analysis methods. The advancement in machine learning (ML) has enabled biologists to discover such patterns on existing biological data and making predictions on unseen data.

ML is a branch of artificial intelligence that enables computers to learn from data and make decisions or predictions without being explicitly programmed. Researchers use ML to build models based on input training data in order to make predictions or decisions. The two main types of machine learning are unsupervised and supervised learning. Unsupervised learning refers to algorithms that learn patterns from untagged data. Without expected ground truth, these algorithms find structures such as clusters and groupings in the data. Supervised learning, on the other hand, involves learning a function that maps the labelled input to output data based on example input-output pairs learned from the training data. In other words, the model learns from historical examples to make predictions or classifications on unseen data.[55] Supervised learning iteratively makes predictions until an accepted level of performance is

reached. Popular supervised learning method including the linear models such as linear and logistic regression and the non-linear methods including support vector machine and ensemble tree-based methods such as random forest and gradient boosting machines.[56] In recent year, a class of ML called neural networks emerges as a popular model and evolved into the field of deep learning. Neural networks are advantageous in handling large and complex data, are flexible in handling unstructured data and does not require manual feature engineering.[57] When choosing the proper ML algorithm for real-world problem, models should be carefully selected based on the structure and complexity of the data (linear vs non-linear), computational resources available, requirement on normalization, and explainability.

Among supervised machine learning algorithms, random forest is a popular and easy-to-use non-linear model. Random forest operates by constructing an ensemble of decision trees during the training process and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It combines multiple decision trees to enhance predictive accuracy and attenuate overfitting bias. Random forest can model complex non-linear functions and performs well in practice comparing to other methods. In genomics research, random Forest is widely used to tackle various challenges, such as gene expression analysis, disease classification, and DNA sequence classification. The ability to handle high-dimensional data, capture complex interactions among genes or features, and good explainability by providing feature importance rankings makes random forest a valuable tool in deciphering the intricacies of large scale genomics data.[58]

Machine learning (ML) has advanced both basic and translational cancer research by enhancing diagnostic accuracy, enabling more accurate patient classification and prognosis prediction using molecular signatures, and discovering disease-associated genetic variant using omics dataset.[59,60] By analyzing vast datasets, ML algorithms have enabled the

development of precise image-based detection systems, allowing for early identification of various cancer types through techniques like deep learning applied to histopathological and radiological images.[60] Furthermore, ML has revolutionized genomic analysis, identifying genetic mutations and biomarkers that predict disease progression and treatment responses, thereby facilitating personalized medicine approaches.[59] In patient classification, ML models integrate clinical and genomic data to stratify patients into specific risk groups, improving treatment decisions and outcome predictions. Additionally, in the realm of prognosis, ML tools have been instrumental in forecasting survival rates and treatment efficacy, using data from electronic health records, imaging, and omics. These applications collectively demonstrate ML's transformative potential in advancing cancer research.

1.4 CUX1 in hematopoietic malignancies

CUX1 is a ubiquitously expressed, non-clustered homeobox containing transcription factor. It is highly conserved evolutionarily and functionally from *Drosophila* to humans. *CUX1* is encoded on chromosome 7q22.1 and it has several RNA and protein isoforms.[61] Historically, CUX1 has been thought to contain three isoforms, the full length p200 and the truncated version p110 and p75, generated by an alternative transcriptional start site or post-translational cleavage, respectively.[62,63] Our lab has shown that p75 is likely a western blot artifact rather than real isoform.[64] This will be elaborated more in detail later in this section. Hematopoietic cells only express the full length p200 isoform,[64] which contains one homeodomain and three CUT repeat DNA-binding domains.[65] Another conserved domain is the coiled-coil domain, which is likely involved in protein-protein interaction. Additionally, the CUX1 N-terminal region has an autoinhibitory domain that likely inhibits DNA binding,

and the C-terminal regions contains two repressive domains that are thought to repress transcription by recruiting histone deacetylases to target promoters.[66,67] (**Figure 1.2**)

As one of the few chromosome 7 genes that is recurrently mutated in cancers, *CUX1* loss has an adverse effect in transformation through either loss-of-function mutation or -7/del(7q). *CUX1* mutation is found in 2-4% of myeloid diseases including MDS, MDS/myeloproliferative neoplasms and AML, and 1~5% in solid tumors.[68,69] Clinically, comparing to MDS and AML patients with wild type *CUX1*, the patients with inactivating *CUX1* mutations have worse survival comparable to patients with -7/del(7q).[38] Additionally, *CUX1* mutations are found in clonal hematopoiesis of indeterminate potential (CHIP), similar as -7/del(7q) in disease development.[70,71] This indicates that *CUX1* mutation could an early driver event in myeloid disease progression. *CUX1* mutation is characterized as mostly monoallelic and fit the mutational signature of TSGs (coding region frameshift and nonsense mutations), suggesting that CUX1 acts as a haploinsufficient TSG.[42,69]

Indeed, there has been strong evidence supporting *CUX1* as a TSG in myeloid malignancies. Knocking-down the ortholog of *CUX1* in *Drosophila melanogaster* leads to hemocyte overgrowth and tumor formation.[49] *CUX1* is often deactivated in human myeloid neoplasms either via -7/del(7q) or recurrent loss-of-function mutations.[49] Through RNA-seq and SNP array analysis in *de novo* AML and therapy related myeloid neoplasms patient samples, McNerney et al. 2013 identified a 2.17 Mb commonly deleted regions on chromosome 7 that contains *CUX1*, and found that *CUX1* is the most differentially expressed gene on this region. In addition, haploinsufficiency of *CUX1* in human hematopoietic progenitor cells led to engraftment advantage in a xenograft model. [49]

Additional functional studies done in the mouse model by our lab has also provided strong evidence for *CUX1* as an important myeloid TSG. Historically, knocking down *CUX1* in mice has been challenging because of several factors. First, *CUX1* is a large gene composed of 33 exons spanning 340 kilobases (**Figure 1.2**). Multiple RNA and protein isoforms further added to the complexity.[61] *CUX1* has two alternative start sites and seven different RNA splicing isoforms.[65] As a result, previous *CUX1* knockdown mice have unintentional alternative spliced *CUX1* RNA that removed the STOP cassette, which led to hypomorphic protein expression and incomplete *CUX1* knockout. Previous models also have undesired extrahematopoietic effects and perinatal lethality.[72–75]

Secondly, *CUX1* shares exons with another gene *CASP* which encodes for a highly expressed Golgi apparatus-associated protein that lacks the *CUX1* DNA binding domains.[76] [77](**Figure 1.2**). *CASP* has so far not been implicated in any human diseases.[78]

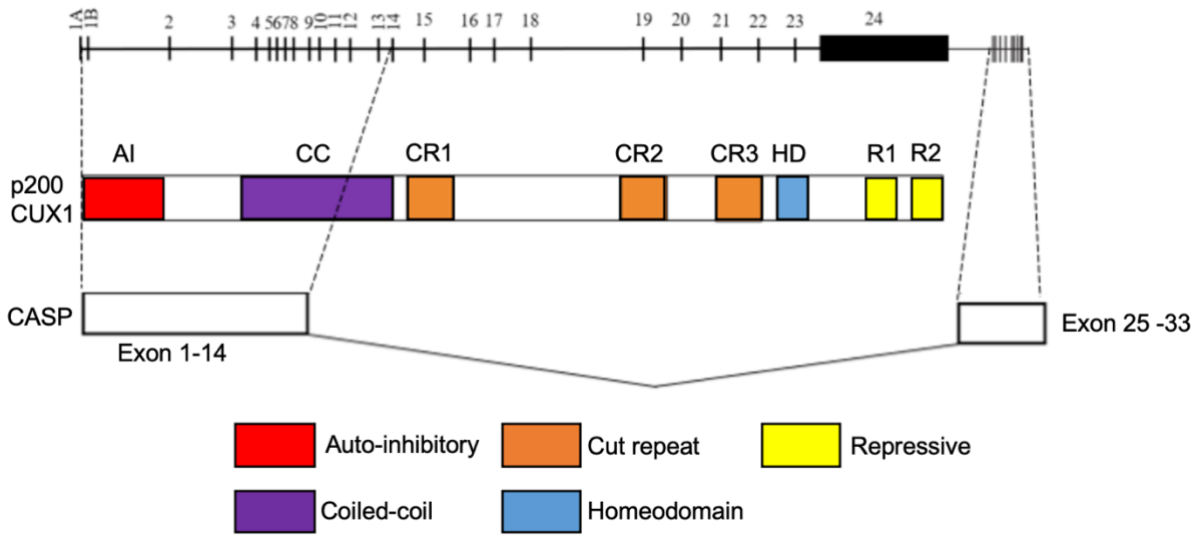


Figure 1.2 Full length CUX1 p200 gene and protein structures, and exons shared with CASP

To overcome these challenges, our lab previously generated two dosage-specific doxycycline inducible shRNA knockdown mouse models that resulted in mid and low *CUX1*

expression in thymocytes (~54% and 12% residual CUX1 protein respectively). The CUX1^{mid} shRNA targets a *CUX1* and *CASP* shared exon and aims to mimic *CUX1* haploinsufficiency, whereas the CUX1^{low} uniquely targets *CUX1* transcripts. CUX1 knockdown leads to dosage-dependent disease phenotype where CUX1^{mid} mice developed normocytic anemia and splenomegaly, and CUX1^{low} mice developed more serious MDS/myeloproliferative neoplasms and anemia with fatal consequences.[79] Taken together, these studies supported *CUX1* to be a dosage-dependent TSG.

Defining the transcriptomic effect of CUX1 from RNA-seq data has also been hampered by the exon sharing between *CUX1* and *CASP*. Standard reference genomes used for transcriptomic analysis do not differentiate *CUX1* and *CASP* transcripts. Transcripts labelled as "*CUX1*" in RefSeq in fact encompass both *CUX1* and *CASP*. This may have masked some CUX1-dependent differentially expressed genes in previous RNA-seq analysis. To address this challenge, I created a customized reference genome where *CUX1* and *CASP*-specific transcripts are differentially labelled in the genome reference annotation file. Details on how this customized reference genome was created can be found in the materials and method section. The new reference genome gives us clearer readout from transcriptomic analysis. RNA-seq analysis showed that human hematopoietic stem and progenitor cells (HSPCs) presented reduced quiescence and increased proliferation, reduced negative regulation of myeloid differentiation, and increased PI3K/AKT/MTOR signals upon *CUX1* loss. Gene set enrichment analysis also showed CUX1-deficient HSPCs showed a concordant gene signature with -7/del(7q) driven MDS patients.[79]

CUX1 also plays critical roles in DNA damage repair. Imgruet et al. 2021 showed that CUX1 recruits the histone methyltransferase EHMT2 to DNA damage sites. The recruitment in turn promotes downstream H3K9 and H3K27 methylation, phosphorylated ATM retention,

gH2AX focus formation and propagation, and, ultimately, 53BP1 recruitment.[80] This series of event eventually lead to DNA damage repair. In the absence of CUX1, the DNA damage response is compromised, and DNA damage is not repaired. CUX1-deficient mice developed clonal hematopoiesis similar to patients post chemotherapy, despite the presence of pervasive unrepaired DNA damage. This ultimately pre-disposed the mice to fatal therapy related myeloid neoplasms.[81] This study provided DNA damage repair as another key mechanism that CUX1 serves as a gatekeeper to prevent myeloid transformation.

Despite the strong evidence supporting CUX1 as a TSG in hematopoietic malignancies, other groups have also reported that CUX1 might serve as an oncogene in some solid tumors and even myeloid leukemia. For example, over-expression of short p75 or p110 CUX1 isoforms in fibroblasts and breast cancer cells have been reported to cause increased proliferation, cell cycle progression, and tumor formation *in vivo.*[63,82,83] p75 CUX1 transgenic mice engendered a higher proportion of adenosquamous mammary carcinomas and lung metastases compared to p110 or p200 transgenic mice.[83] Transgenic mice overexpressing p75 isoform developed myeloproliferative disease-like myeloid leukemia.[84] Krishnan et. al. 2022 found that HSPCs only express the full p200 isoform and p75 isoform is likely an artifact from the denaturing condition of western blot. Both RNA and protein of p75 CUX1 are not detected in human AML and breast cancer cell lines, which previously thought to express p75.[64] Furthermore, by integrating epigenome data form public database encompassing more tissue and cell types, no active transcription start sites (TSS) and promoter-specific epigenetic marks were detected in the predicted p75 promoter region.[64] Because p75 binds DNA more stable than p200 *CUX1,* and most studies that reported the oncogenic effect of *CUX1* rely on over-expression models using cDNA, the results could be caused by the artificial overexpression of p75 which stoichiometrically interferes with the

endogenous CUX1 protein or blocks the full length p200 from binding to target genes. These interactions could disrupt the p200 CUX1 tumor suppressive function. Although 7q copy number gains and *CUX1* overexpression has been documented in cancer cell lines,[65] no casual-effect studies have been performed to define oncogenic role of *CUX1* under endogenous contexts. *CUX1* might merely be a passenger in amplification events because there are major oncogenes including *EGFR*, *BRAF*, *CDK6*, and *EZH2* located also on 7q. Taken together, the growing body of evidence summarized earlier in this section supports *CUX1* as a TSG under most cancer context, and the contradictory oncogenic role of *CUX1* observed might be due to the limitation of overexpression system and passenger effects.

In summary, the evidence accrued in large-scale cancer genome re-sequencing, *in vitro*, and *in vivo* studies provided strong evidence to support *CUX1* as a vital tumor suppressor. Collectively, the research so far showed that the ways *CUX1* functions is complicated and multifaceted. Unravelling the context-dependent role of *CUX1* in driving oncogenesis is crucial to inform novel therapeutic strategies utilizing *CUX1* as a drug target or biomarker. Notably, since myeloid malignancies are clonal disorders fundamentally driven by aberrant cell fate commitment, understanding the role of *CUX1* in dictating hematopoietic stem cell fate will provide another key mechanism in cancer development.

1.5 Epigenetic regulation of stem cell fate

Multipotent tissue resident stem cells are essential for the generation, maintenance, and function of adult tissues. Defects in stem cell homeostasis and lineage commitment underly myriad human diseases, including cancer.[85] Over the entire lifespan of stem cells and their immediate progeny called progenitor cells, they face multiple types of fate choices including self-renewal, differentiation, and commitment to mature cell fates. The mechanisms governing lineage determination are incompletely understood and remain a fundamental

question in developmental biology. Insights into the process of what elements exist in the cell fate determination process, and how they regulate stem cell fate is central to develop novel therapeutic interventions for diseases caused by stem cell dysfunction.

Transcription Factors (TF)

TFs are DNA-binding proteins that regulate the transcription process from DNA to RNA.[86] They function as key regulators for development patterning, cell fate and control specific signalling pathways.[87] TF protein sequence, regulatory regions and functions are highly conserved across species, implying conservation of gene expression regulatory networks across species.[88] Mechanistically, TFs bind to gene regulatory elements such as promoters and enhancers, and can subsequently either impede or promote DNA transcription of the target genes via the recruitment of proteins that physically remodel nucleosomes, enzymatically modify histones and DNA, and regulate RNA polymerase machinery.[86]

The location preference of TF binding is determined by DNA sequences called "motifs". Motifs are short DNA sequences (typically 6 – 12 base pairs) preferred by each individual TFs for binding. Although motifs are typically enriched in the corresponding TF binding sites determined by experiments such as chromatin immunoprecipitation sequencing (ChIP-seq), there is often only partial overlap between the motif sites and actual TF binding sites. Furthermore, a typical metazoan gene body usually contains multiple sites for a TF to bind.[89] These facts collectively imply motif redundancy and non-specificity in TF-motif matching. Studies over the years have observed that most metazoan TFs work together to achieve the needed specificity in DNA binding and regulatory functions.[90] However, the molecular mechanism of how different TFs interact with each other to achieve desired gene expression regulation remains poorly understood.

Transcription factors are the master regulators of cell fate during the development process. While cell fate decisions are influenced by extrinsic factors, such as cell-cell signalling and growth factors, intrinsic factors, namely, epigenetic regulators and TFs are ultimately responsible for integrating these cues to guide the genomic reprogramming required for cell-type specific gene expression.[91] Some TFs are ubiquitously expressed across different tissue types. These TFs are often involved in regulating a wide array of fundamental processes needed across tissue types. Example includes *CTCF*, which facilitates regulatory sequence interaction by creating boundaries between 3D chromosome topological associated domains,[92] and *SP1*, which regulates a wide array of cell survival and proliferation genes.[93,94] Roughly one thirds of the known human TFs are expressed in a tissue-specific manner.[86] These TFs typically regulate differentiation or other specific functions of the tissue types. Loss of these TF frequently lead to differentiation/development block. For example, *SCL/TAL1* is expressed in the hematopoietic system to regulate early HSC differentiation.[95] *GATA1* is expressed exclusively in the erythroid lineage to promote erythropoiesis.[96–99] *PU.1*, *CEBP/a* and *GFI-1*, on the other hand, are expressed in the myeloid lineage and promotes differentiation of various myeloid cell maturation.[100–102] It is noteworthy to mention that despite their universal expression pattern, ubiquitous TFs can also regulate lineage-specific gene expression by interacting with tissue-specific TFs or proteins, or be post-translationally modified in a tissue specific manner. An example is the universally expressed TF *OCT1*, which could regulate lymphoid cell differentiation by interacting with B and T cell specific proteins OBF1 and IL-3.[103,104]

Cell type-specific enhancers

Among the regulatory elements TFs bind to, enhancers have major roles in determining cell fates. Different expression combinations and dosages of genes dictates what

fates a stem cell differentiate into. The human body contains hundreds of different cell types and all of them share the same primary DNA sequences.[105] Enhancers are non-protein-coding DNA sequences that serve as substrates for TF binding and regulate cell-type specific gene expression. Enhancers can locate anywhere relative to their target gene, including up- or downstream, and within introns.[106] Extensive research over the years have shown that enhancers not necessarily regulate the closest gene, but they can exert regulatory function to faraway loci through 3D chromatin looping.[107] Furthermore, one enhancer could regulate multiple genes and a single gene could be regulated by multiple enhancers.[108]

Enhancer states are categorized into inactive, primed, poised, or active enhancers.[109] Inactive enhancers frequently locate in compact chromatin regions and are devoid of TF binding. Primed enhancers are bound by sequence-specific TFs, and the DNA accessibility is established by such bindings. They are characterized by H3K4me1 deposition. However, primed enhancers require additional events such as recruitment of additional TFs, co-activators, and active histone modification H3K27ac to become active. Poised enhancers are primed enhancers that also contain repressive chromatin marks such as H3K27me3. They are mostly found in stem cells such as embryonic stem cells. In response to various signalling cues, poised enhancers often become active by TF binding, which recruits chromatin remodeler complex to displace nucleosome and further opening the DNA, as well as enzymes such as histone demethylase (HDM) to remove H3K27me3 and histone acetyltransferases (HAT) to deposit the activating H3K27ac marks. Co-activators and transcriptional machinery components such as P300-CREB-binding protein (CBP), the mediator complex and RNA Pol II are subsequently recruited to initiate the transcription process.[105]

Although there are millions of enhancers in the human genome, only a small percentage of enhancers becomes active in each cell type. These cell-type specific enhancers

are ultimately responsible for regulating the expression of sets of genes whose expression leads to distinct cell fates. The selection and activation process of these cell-type specific enhancers typically involved the binding of a class of TFs called "pioneer factors".

Pioneer factors (PF)

Most TFs bind open chromatin regions, but a subset of TFs, termed "pioneer factors" can bind closed nucleosomal DNA ("heterochromatin"), recruit other non-pioneer TFs and chromatin remodeler proteins to promote *de novo* DNA accessibility at these sites.[110] Cells only use a tiny fraction of the genetic information to translate genes to proteins at any given time. The rest of the genetic sequences are hidden behind and packaged away by the nucleosomes. PFs function as the "opener" that make these hidden genetic elements accessible. The first evidence of pioneer TF is *FOXA*, which was shown to bind target sequence that wrapped around and are thus occluded by nucleosomes.[111] Stable cell fates rely on mechanisms that maintain DNA accessibilities to TFs, and pioneer factors could alter this state and establish new cell fate by opening otherwise closed DNA regions for DNA and histone modifications that mask or unravel genomic regions through chromatin remodelling. This opening action by PFs allows new TFs to bind and initiate subsequent recruitment of other transcriptional machineries, which ultimately result in cell-type specific gene expression programs. Thus, pioneer factors are the master regulator in development and stem cell differentiation.[112]

Given the central role in regulating cell fate, aberrant expression due to mutations and aneuploidy of PFs could lead to serious consequences in cancer. For example, mutations in the PF *FOXA1* were shown to promote prostate cancer by altering chromatin landscape that perturbed normal luminal epithelial differentiation process.[113] PFs can enable estrogen and

androgen receptors to bind chromatin and promote oncogenesis in hormone-dependent breast cancer.[114] Overexpression of the PF *HOXA9* in myeloid and B lineage progenitor cells activates leukemia-specific *de novo* enhancer and promote leukemogenesis.[115] In addition, PFs can also fuse with each other and form oncogenic fusion proteins.[116] Therefore, pioneer factors are promising targets for novel therapeutic development.

The definition of pioneer factor is actively evolving. According to the classical model, a pioneer factor must act in a sequential fashion to recognize its binding motifs in heterochromatin regions, and subsequentially recruit cooperative TFs and other components such as chromatin remodelers to activate the target genes.[117–119] The underlying assumption of this classical definition is that PFs should be able to bind to majority of the target motifs and non-PFs can only bind to open chromatin. Hansen et. al. 2022 challenged this binary definition of PFs. They ectopically expressed the classical endodermal PF *FOXA1* and non-PF *HNF4* in K562 cells and found that instead of following the conventional "two step" process where *FOXA1* binds and opens inaccessible chromatin regions and *HNF4* following suit, both factors can access and bind heterochromatin and pioneer for each other. They further find that the main difference is that the DNA binding of the classic PF *FOXA1* does require fewer copies of the recognition motifs than HNF4.[120] The authors argued that instead of categorizing TFs into the binary PF and non-PFs, TFs processes a spectrum of pioneer activities categorized by properties such as the binding motif affinity. In addition, TFs can switch between PF and non-PF modes depending on expression dosages. A recent study showed that *SOX2* loses its pioneering activity and switch to non-PF mode and collaborate with other TFs to bind open regions when expressed at lower levels.[121] Based on these studies, when defining the pioneering role of a TFs, it is more accurate to use the terminology "TF with pioneer activity" than "pioneer factor".

Pioneer factors play a central role in hematopoiesis. Hematopoiesis cell fate determination is a complex and closely orchestrated process that involved multiple TFs, many of which are lineage restricted.[122] Research over the years has found several TFs with pioneer activities. For example, *EBF1* specify the lymphoid progenitor cells towards B cell fate by collaborating with *PAX5*.[123] *PU.1* steers HSPCs towards myeloid and macrophage fates by collaborating with another pioneer factor *C/EBPa*.[124] *RUNX1* is another PF that is required for maintaining hematopoietic stem cell homeostasis and multilineage differentiation.[125,126] In the T cell lineage, *TCF1* is required to open enhancers that establish T cell identity.[127] Recent studies have also shown *KLF1* might function as a pioneer factor to recruit *GATA1* and *SCL* and promote erythropoiesis.[128] However, our understanding of PFs in hematopoiesis is still very limited. Notably, we do not know whether there are apex PFs functioning in the very early stage of hematopoiesis to steer HSC towards different fates.

Chromatin remodelers and the BAF complex

The position and density of nucleosome regulate the accessibility of binding sites to TF and the transcription machinery.[129] Therefore, nucleosome positional, phasing and density must be finely regulated to enable transcription to happen at the right genomic loci and at the right time. There are two classes of chromatin remodelling proteins that regulate the accessibility to nucleosomal DNA. First, there are histone-modifying complexes that covalently modify histones by depositing or removing histone marks. These modifications include methylation, acetylation, phosphorylation, sumoylation and ubiquitination etc.[130] These processes alter the binding affinity of histones and DNA strands, thus loosening or tightening the condensed DNA wrapped around histones.[131] The other class of chromatin modifying proteins are called ATP-dependent chromatin remodelers. They all have an ATPase subunit and rely on the energy provided by ATP hydrolysis to move along the target

nucleosome to alter chromatin accessibility by repositioning, ejecting, or evicting nucleosomes.[132] There are four categories of chromatin remodelers, including the imitation switch (ISWI), switch/sucrose non-fermentable (SWI/SNF, or BAF), chromodomain helicase DNA-binding (CHD), and INOsitol requiring 80 (INO80). Functionally, INO80 regulates DNA transcription and repair and is responsible for removing and replacing histones by canonical or related variants.[133] ISWI and CHD are associated with modulating nucleosome organization following DNA replication by regulating the mobilization of nucleosome and the length/position of nucleosomal linker spacings.[134]

The BAF complex chromatin remodeler

The BAF (BRG1/BRM-associated factor, or SWI/SNF) complex is a large multi-subunit protein complex belonging to ATP-dependent chromatin remodelers. The BAF complex is the key regulator of nucleosome positioning and is responsible for sliding through the nucleosome to regulate the nucleosome spacing, density and phasing.[135] The BAF complex frequently localizes to enhancers when recruited by lineage-specific TFs. At these enhancer sites, the BAF complex modulates DNA accessibility which is required for activating gene expression.[136] Mammalian BAF complex is composed of 10-13 subunits and contains three subfamilies based on subunit composition: canonical BAF (cBAF), polybromo-associated BAF (PBAF) and the recently discovered non-canonical BAF (ncBAF).[137] All three subfamilies contain the mutually exclusive catalytic ATPase subunits *SMARCA2* or *SMARCA4*, which generate energy by hydrolyzing ATP. The energy generated enabled the BAF complex to remodel chromatin through nucleosomal sliding and eviction. All BAF subfamily complexes share many common subunits including *SMARCC1*, *SMARCC2* and *SMARCD*, but also contains variable subunits and confer each subfamily unique functions.[138] cBAF activity is strongest at enhancers, while PBAF and ncBAF are

enriched at promoters, with some degree of enhancer binding. Understanding on the distinction of the three subfamilies function is still very limited and is an active research area. Functionally, except modulating enhancer accessibility and cell fate-specific transcription, the BAF complexes have essential roles in DNA damage repair.[139] For example, cBAF and PBAF are involved in both homologous recombination and non-homologous end joining repair processes.[140,141] SMARCA4 and the cBAF specific *ARID1A* have been shown to be recruited to DNA damage sites and help with DNA repair.[142,143] *SMARCA4* have been shown to promote DNA accessibility at DNA damage sites and collaborate with *PARP1* to initiate DNA repair machineries, as well as by inducing histone H2AX phosphorylation[144,145] These studies showed that the BAF complex is essential for genome integrity maintenance, and dysfunction of the complex could lead to mutations that further drives cancer development.

The BAF complex is highly mutated across different cancer types. Mutations encoding the BAF complex subunits collectively occur in ~25% of all cancers.[146] Studies elucidating oncogenic roles of BAF mutations only emerged fairly recently. In 1998, *SMARCB1* biallelic mutations were identified in a rare pediatric soft tissue sarcoma rhabdoid tumors.[147] Subsequent animal studies showed that that knocked down *SMARCB1* led to highly penetrant cancer predisposition with 100% of the mouse developing T cell lymphoma and rhabdoid cancer within 11 weeks.[148] With the advent of next generation sequencing technology, researchers discovered that multiple BAF subunits are widely mutated in cancer. For example, *PBRM1* is mutated in >40% of clear cell renal cell carcinoma.[149] *ARID1A* and *SMARCA4* are frequently mutated in multiple cancer types.[150–152] *SS18* subunit fusion to *SSX* form an oncoprotein, which is a driver in synovial sarcoma.[152]. A common observation is that different subunits of the BAF complex are mutated selectively in different cancer types, and cancer cells harboring specific BAF subunit mutations developed dependency on the

functionally-related paralogs of specific subunits.[153] This had led to various therapeutic development efforts to target BAF subunits using the logic of synthetic lethality. For example, genome-wide shRNA and CRISPR screens revealed that cancer cells lines harboring *ARID1A* mutations rely on the compensatory roles of *ARID1B* to maintain growth. ARID1B inhibitors are thus being tested to target *ARID1A* mutated cancer.[154,155] With a deeper understanding of the oncogenic mechanism of different BAF subunits under different context, there will be more promising therapeutic strategies in this space.

As BAF proteins lack intrinsic DNA binding domains, they depend on TFs for DNA targeting specificity. In hematopoiesis, various TFs have been shown to modulate gene expression by recruiting the BAF complex and remodelling the chromatin. In the myeloid lineage, the pioneer factor *PU.1* recruits the BAF complex to access and remodel chromatin *de novo*, and promote accessibility at enhancers co-bound by collaborative TFs.[156,157] Inhibition of the BAF complex redistributed *PU.1* to promoters and induces leukemic differentiation-related gene expression.[157] *RUNX1* is a PF important in promoting myeloid lineage differentiation. It recruits the BAF complex to control target gene expression.[158] The myeloid differentiation PF and TSG *C/EBPα* also requires the recruitment of the BAF complex to promote myeloid -specific gene expression and inhibit cellular proliferation.[159,160] In the erythroid lineage, *GATA1* recruits the BAF complex to the β-globin locus control region. BAF then mediates *GATA1*-dependent chromatin looping and transcriptional activation.[161,162] A recent study also showed that the nuclear factor hemogen interacts with *GATA1* and helps recruit the BAF complex in order to form the activating LDB1 complex and promote pro-erythroid differentiation gene expression during terminal erythropoiesis.[163] The erythroid pioneer factor *EKLF/KLF1* was shown to collaborate with the BAF complex to promote accessible β-globin promoter and promote the stage-specific expression of human β-

globin gene.[164,165] Nonetheless, these TFs only account for a fraction of BAF chromatin binding, implicating additional, yet unknown, hematopoietic pioneer TFs.

1.6 CUX1 in hematopoietic stem cell fate

CUX1 is a conserved important regulator of cell differentiation in multiple tissue systems. In *Drosophila Melanogaster*, different expression levels of *CUX1* ortholog *Cut* were found to regulate dendrite morphology in dendritic arborization sensory neurons. Loss of *Cut* reduces dendrite growth and class-specific terminal branching, whereas the ectopic overexpression of *Cut* in lower-level neurons leads to transformation of branch morphology to that similar in high-Cut neurons.[166] This observation is also consistent in mice, where *Cux1* and its paralog *Cux2* regulate dendrite branching, spine development and synapse formation in neurons of the cerebral cortex.[167] The regulatory role of *CUX1* in tissue development goes beyond the nervous system. Ellis et al. 2001 mutated *Cux1* in mice by replacing the C-terminal Cut repeat 3 and homeodomain exons with an in-frame lacZ gene by targeted mutagenesis. They found that the mice on inbred genetic background where both *Cux1* alleles were mutated died shortly after birth due to retarded differentiation in lung epithelia, while the more genetically diverse outbred background mice with only one *Cux1* allele mutated experienced less fatal phenotypes less lung development damage, had longer survival than the inbred mice, but had severely damaged hair follicle development.[75] Collectively, these studies buttress the essential role of *CUX1* in development process of multiple tissue systems.

CUX1 is a key cell fate regulator in hematopoiesis. Using the shRNA knockdown mouse model, our lab previously reported that *CUX1* regulates HSPC homeostasis and differentiation.[79] Specifically, *CUX1* maintains HSC quiescence and repress proliferation.

Knocking down of *Cux1* to both Cux1^{mid} and Cux1^{low} levels leads to increased HSPC proliferation but decreased long-term self-renewal capability, indicating loss of *Cux1* resulted in stem cell exhaustion. Furthermore, knocking down *Cux1* also led to a decrease of long term-HSC (LT-HSC) population fraction in the quiescent G_0 state and increase of the proliferative G_2/S phase populations. Gene set enrichment analysis (GSEA) using RNA-seq on bulk primary human HSPC showed that knocking out *CUX1* lead to downregulated stem cell quiescence gene signature and upregulated proliferative signature, concordant with the *in vivo* experimental observation. *CUX1* knockdown induced upregulation of gene regulated by PI3K signalling, which is notable because increased PI3K signalling is associated with HSC exit from quiescence and proliferation.[79] In addition, Cux1^{low} bone marrow cells have increased PI3K substrate AKT phosphorylation and decreased expression of the PI3K inhibitor *Pik3ip1,* consistent with previous report that CUX1/Cut suppress PI3K activities.[69,168] In addition to regulating HSC homeostasis, *CUX1* also promotes healthy erythropoiesis. Loss of *Cux1* lead to decreased red blood cell count and anemia, which is especially manifested in the Cux1^{low} condition. Mechanistically, functional assays showed a differentiation block in the orthochromatophilic erythroblast stage contributes to the impaired erythropoiesis.[79] Notably, loss of *CUX1* caused the expansion of white blood cell population including monocytes and granulocytes, and this increase is also manifested earlier in the differentiation trajectory as increased common myeloid progenitors (CMP) and granulocyte monocyte progenitors (GMP) population.[79] In HSPC RNA-seq, GSEA analysis orthogonally showed that the "Negative regulation of myeloid cell differentiation" GO term was downregulated after knocking down *CUX1*, consistent with the *in vivo* experiment observation and indicates that *CUX1* represses myeloid differentiation. Taken together, our lab showed that *CUX1* knockdown promotes PI3K signalling, drives HSC exit from quiescence and proliferation, results in HSC exhaustion, and impairs erythropoiesis at the

expanse of myeloid expansion. The indispensable roles of *CUX1* in hematopoietic cell fate indicate that *CUX1* exerts tumor suppressor activity via transcriptional regulation of HSPC functions, yet the mechanisms by which CUX1 coordinates gene expression remains unclear.

Understanding the epigenetic mechanism of *CUX1* function is the key to elucidate how it regulates gene expression. In hematopoietic cells, *CUX1* genomic binding features are distinct, as revealed by ChIP-seq analysis in three human cancer cell lines including the K562 cells.[169] It exhibits a preference for distal enhancers over promoters and co-occupying sites with RNA polymerase II, EP300. This indicate *CUX1* pays a central role in the cis-regulation of transcription. In addition, *CUX1* also co-localizes with cohesin and CUX1 binding sites are enriched at DNA looping contact points, indicating CUX1 regulates genes via looping cis-regulatory elements to promoters.[169] RNA-seq analysis following *CUX1*-knockdown uncovered the dual role of *CUX1* as an activator and repressor.[79,169] The regulation of pathways involving cell cycle progression, proliferation, apoptosis, multilineage differentiation, and quiescence further emphasizes its multifaceted functions. *CUX1* preferential binding to distal enhancers and its unique analog model of dose-sensitive gene regulation hint at disparate roles in different cell types and possibly under different stress conditions.[169] In summary, *CUX1* acts to regulates gene expression through distal enhancers that loop to target promoters.

CUX1 itself can bind closed nucleosome, indicating pioneer activity. Using *in vitro* electrophoretic mobility shift assays and DNase I footprinting experiments, Last et. al. 1999 showed that CDP/cut can bind to its recognition motifs on nucleosome cores reconstituted from histone H4 gene promoter (-90 to +75).[170] While *CUX1* binding destabilizes the nucleosome,[171] *CUX1* binding alone does not cause nucleosome displacement. This suggested that there are other cooperative co-factors/chromatin remodelers necessary for *CUX1* to

remodel chromatin. Taking together, research to date showed that *CUX1* is important in regulating lineage determination and transcription, binds mostly to distal enhancer regions and displays ability to bind to closed nucleosome regions. All these features are consistent with those of a PF. Therefore, a major part of my work sets to test the hypothesis whether *CUX1* function to regulate cell fate as a PF, and if so, what the molecular mechanisms are.

1.7 Erythropoiesis

Erythropoiesis, a critical component of hematopoiesis, is the maturation processes that produce mature red blood cells (RBCs). These cells are indispensable for the transport of oxygen to various tissues and organs, a function that is essential from embryonic development to adulthood and throughout the entire lifespan. Red blood cells represent the terminal differentiation state within an intricate hierarchy that originates from HSCs.[172] Early erythroid progenitors advance through a sequence of meticulously orchestrated maturation stages. The entire erythropoietic process is tightly regulated to synchronize RBC production with the physiological oxygen demands of the body. In healthy human adults, the bone marrow produces approximately 2×10^{11} new erythrocytes each day.[173] The oxygen-carrying capability of RBCs is mediated by hemoglobin (Hb), a complex protein within RBCs. Hemoglobin biosynthesis is regulated by two distinct multi-gene clusters located on human chromosomes 16 (α globin) and 11 (β globin). Hemoglobin composition undergoes a developmental transition from fetal to adult form. In the fetal stage, two γ genes pair with α-globin genes to form Hb F ($\alpha_2\gamma_2$). Postnatally, the α globin product associates with the β globin product to produce Hb A ($\alpha_2\beta_2$), the predominant adult hemoglobin variant. The switch from fetal to adult hemoglobin initiates prior to birth and is predominantly completed by the age of 6 months.[174]

Erythropoiesis initiates within the bone marrow through a tightly regulated stepwise process. Pluripotent progenitor cells embark on a differentiation journey, first transforming into immature erythroid progenitors. This early phase of erythropoiesis encompasses the formation of primitive burst-forming unit erythroid cells (BFU-E) and the subsequent development into the more mature colony-forming unit erythroid cells (CFU-E). Hematopoietic cytokines, notably stem cell factor (SCF) and interleukin-3 (IL-3), play a crucial role in orchestrating this stage of erythropoiesis.[175] Progressing through erythropoiesis, cells enter terminal erythropoiesis stage and evolve into proerythroblasts (ProE), basophilic erythroblasts (Baso), polychromatic erythroblasts (PolyE), and orthochromatic erythroblasts (OrthoE). These stages the lead to the formation of reticulocytes, which mature into functional red blood cells.[176] The transformation, termed reticulation, occurs as the cells migrate from the bone marrow to the bloodstream. During this transit, erythroid cells undergo a remarkable transformation, shedding their nuclei and most organelles, ultimately acquiring the distinctive biconcave shape characteristic of mature RBCs.

A complex network of TFs, cytokines and hormones orchestrate the whole process of erythropoiesis.[177,178] Central to this process are transcription factors such as GATA-1, TAL1/SCL, EKLF, and NF-E2, which play crucial roles in different stages of erythroid differentiation, from progenitor commitment to terminal maturation.[179] Among them, GATA1 is the key regulator, and is responsible for early erythroid lineage commitment, differentiation, and survival.[180,181] The roles of GATA1 in erythropoiesis will be discussed in detail in the next section. The hormone erythropoietin (EPO) signalling pathway is also pivotal, with EPO receptors activating specific kinases upon binding to EPO, the principal regulator of erythropoiesis. Produced primarily by the kidneys, EPO production is regulated

by an oxygen-sensitive feedback loop and is activated by hypoxia.[182] In response to hypoxia, or reduced oxygen availability, EPO production increases, stimulating erythropoiesis and RBC production. Conversely, when oxygen levels are sufficient, EPO production diminishes, preventing excessive RBC production.[183] Additionally, the PI3K/AKT pathway has been identified as a significant regulator, impacting cell survival, differentiation, and the prevention of apoptosis during the maturation of red blood cells.[184] Furthermore, hypoxia-inducible factors (HIFs) play a master regulatory role, especially under hypoxic conditions, modulating erythropoiesis by controlling EPO expression and coordinating with various other proteins involved in iron metabolism.[185] These intricate networks of transcription factors and signalling pathways ensure the precise regulation of erythropoiesis, ensuring the continuous production of red blood cells.

The dysregulation of erythropoiesis causes a broad spectrum of human diseases, such as anemia, β-thalassemia, sickle cell disease, and polycythemia vera.[176] Additionally, several hematological malignancies are directly attributable to aberrant erythropoiesis. Notably, MDS and erythroleukemia are caused by an impaired differentiation in the initial stages of erythropoiesis, leading to conditions such as red cell dysplasia and erythroid bone marrow hyperplasia, as well as the abnormal programmed cell death of erythroid progenitor cells.[186] A comprehensive understanding of the molecular mechanisms governing erythropoiesis is essential, as it holds the potential to unlock new pathways for therapeutic intervention and innovation.

1.8 GATA1 in hematopoiesis

GATA-binding factor 1 (*GATA1*) is the founding member of the GATA family TFs and is a central regulator in the intricate regulatory network governing erythropoiesis. It functions as a gene expression activator or repressor depending on contexts.[187] *GATA1* was first discovered in 1988 as a β-globin enhancer- and promoter-binding factor.[188,189] Subsequently, genetic studies in zebrafish and mouse established the central role of *GATA1* in promoting erythropoiesis.[96–99] Pronounced phenotype upon *Gata1* depletion was observed in mice, where erythroid progenitor maturation arrest resulted in ablation of primitive and definitive erythropoiesis and embryonic lethality.[98,99] *GATA1* is expressed in and important for differentiation of erythrocytes, megakaryocytes, mast, eosinophil. basophil and dendritic cells.[190–195] With co-regulators such as friend of *GATA1* (*FOG-1*), *GATA1* regulates gene expression in key pathways including heme-biosynthesis, cell cycle, proliferation, and apoptosis (PMID: 28179282). *GATA1* binds to specific DNA sequences known as (W)GATA(R) motifs through conserved dual zinc finger domains.[180] Despite the fact that distribution of GATA motifs is very promiscuous and there are millions of GATA motifs genome-wide.[196] GATA target genes represent a very small fraction of loci containing GATA binding motifs, with *GATA1* only binds to <1% of the GATA motif, as an example.[196,197] Following the discovery of *GATA1*, other GATA family TFs with overlapping and unique functions have been discovered and cloned. *GATA1-3* are mainly expressed in the hematopoietic system, and *GATA4-6* are expressed in other tissue systems including intestine, lung, and heart.[198,199]

GATA1 is not the sole member of the GATA family involved in hematopoiesis. *GATA2*, another GATA factor, plays a distinct but complementary role in early hematopoietic stem cell maintenance and early erythroid development.[200] *GATA2* shares similar binding

sequence preference with *GATA1*, and is highly expressed in the early HSPCs, mostly preceding but also overlapping the expression of GATA1. Deletion of *Gata2* in mice led to embryonic lethality due to broad collapse of hematopoiesis, showing the indispensable role of *Gata2* in regulating HSPC homeostasis.[201] In erythroid differentiation, the "GATA switch" phenomenon happens where *GATA1* competitively replaces *GATA2* chromatin binding as stem cells commit to the erythroid fate. This concordantly involves the decrease in *GATA2* and increase in *GATA1* expression levels.[202] A genome-wide analysis of enhancer usage showed that 30% of GATA-bound enhancers underwent GATA switch, which manifested in altered transcriptional outputs that drives healthy erythropoiesis.[203] Mechanistically, the GATA switch is regulated through inhibitory autoregulation. *GATA1* directly represses *GATA2* transcription by displacing it from chromatin sites. In erythroblasts, the co-factor *FOG1* recruits the chromatin remodeler NuRD and promotes the *GATA1* binding at an ~ 70 Kb upstream autoregulation region occupied by *GATA2* itself.[202] This binding event leads to direct inhibition of *GATA2* gene transcription.

The interaction of GATA1 with other hematopoietic TFs is also vital in orchestrating proper differentiation. By adjacent E-Box motifs, GATA1 co-bind DNA and collaborate with TAL1/SCL to form a complex and recruit other non-DNA binding proteins such as LMO2 and LDB1 to regulate expression of erythroid-specific genes.[204] GATA1 also cooperate with KLF1 by co-occupying genomic loci to promote erythroid specific gene expression.[205] In addition, GATA1 also recruits the BAF complex to mediate the GATA1-depedent chromatin looping and transcriptional activation of α- and β-globin loci.[165,206] Besides collaborating with other hematopoietic TFs and chromatin remodelers, GATA1 can also form antagonistic relationship with PU.1 which is a TF that promotes myeloid differentiation. The two proteins physically interact and antagonize each other's actions.[207] Inhibiting *GATA1* expression shifts

HSPC fates towards myeloid fate, while inhibiting PU.1 expression leads to a shift to erythroid fate.[207,208]

By performing ChIP-seq and CUT&RUN, we found that CUX1 and GATA1 share many binding targets genome-wide in both K562 and primary human HSPCs (**Figure 4.1A**). Given the fact that CUX1 promotes healthy erythropoiesis,[79] we hypothesize that CUX1 might coordinate with GATA1 in regulating erythropoiesis. Therefore, elucidating the exact molecular mechanism of how CUX1 and GATA1 will provide exciting new insights on erythropoiesis mechanisms.

1.9 Aims of this thesis: critical gaps in knowledge that will be addressed.

The thesis aims to elucidate the molecular mechanisms by which CUX1 influences the fate of hematopoietic stem and progenitor cells (HSPCs). Despite existing evidence on role of CUX1 in promoting HSC homeostasis, regulating erythroid vs myeloid lineage balance, as well as its regulatory role in various tissue stem cell fate in *Drosophila melanogaster* and mice,[69,79,166] the molecular mechanisms by which CUX1 determines cell fate are poorly understood. The central objective is to understand how CUX1 regulates stem cell fates, a question that remains largely at the phenotypic level.

The thesis is structured into three distinct yet interconnected parts, each targeting a unique aspect of CUX1's influence on cell fate and cancer development:

Part I: I developed a supervised machine learning classifier to predict tumor suppressor genes on chromosome 7. Given the prevalence of -7/del(7q) aneuploidy in various cancers, especially myeloid malignancies, this part filled the gap for a systemic approach to

identify and characterize tumor suppressor genes (TSGs) on chromosome 7. A main bottleneck is the sheer number of genes on chromosome 7 makes it unpractical to discover TSG one-by-one. With the advance in next generation sequencing, high throughput screens, and gene editing technology, there have been abundant published genome-wide perturbation screens using hematopoietic cancer cell lines. Supervised machine learning approach is ideal to mine these rich data, learn the behaviors of the canonically known TSGs, and predict the TSG-likeness for all chromosome 7 genes. The result will be a ranked list of all human chromosome 7 genes with high-to-low TSG likeness scores. Such a list will help to validate in vitro screens looking for TSGs, and serve as a starting reference for future experiment aiming to further investigate individual TSGs.

Part II focuses on dissecting the molecular mechanisms through which CUX1 dictates HSPC fate. Specifically, this part tests the hypothesis that CUX1 functions as a PF in determining stem cell fate. Pioneer TFs are master regulators of cell fate by opening enhancer elements and subsequently recruiting other TFs and co-activators to initiate lineage-specific gene expression programs. A key gap in our knowledge is there are only a few known PFs in the hematopoietic system, and none of them reside in the hematopoietic apex. The mechanism of cell fate determination at this early stage remains cryptic. Besides promoting HSC homeostasis and erythroid vs myeloid cell fate choice, research from our lab and others have shown that CUX1 display preliminary biochemical and epigenetic properties similar to pioneer factors, including binding nucleosomal DNA *in vitro*, bind to distal enhancers, co-bind with transcriptional machinery factors and proteins responsible for long-distance looping. My research aims to use unbiased proteomics and genome-wide epigenome assays to determine the interaction protein partners of CUX1, CUX1 genome-wide binding pattern

in primary HSPCs, effect of CUX1 in chromatin accessibility and the resulting gene expression consequences.

Part III delves into the interactions between TFs, specifically between CUX1 and GATA1, both of which promote healthy erythropoiesis. TFs interact with each other through many different modes to coordinate gene expression. These modes include co-binding, recruitment, forming a co-regulatory complex and antagonizing each other physically or transcriptionally. A key gap in knowledge remains whether CUX1 and GATA1 collaborate/interact with each other to regulate erythropoiesis. A key observation from part II is that CUX1 binding sites in K562 and HSPCs are enriched for GATA motifs. ChIP-seq and CUT&RUN analysis in these two cell systems further showed that CUX1 and GATA1 share many common binding sites. Based on these preliminary data, this part of my thesis aims to elucidate the molecular mechanism of if and how CUX1 engage with GATA1 in hematopoiesis. By exploring their co-binding patterns and the effects of CUX1 loss on GATA1 occupancy and function, this part aims to uncover the collaborative mechanisms that regulate erythropoiesis. The discovery will add to part II as a more complete picture of CUX1 as a hematopoietic PF. This could shed light on the phenotypes observed in myeloid cancer patients with CUX1 mutations or with -7/del(7q) abnormalities, potentially guiding the development of new treatments.

Overall, the thesis proposes to offer mechanistic insights into myeloid malignancies associated with -7/del(7q), providing a comprehensive tool for identifying TSGs and clarifying the role of CUX1 as a pioneer factor in HSPC fate determination. It seeks to unravel the complex interactions between two major erythropoiesis regulators GATA1 and CUX1, which could offer key mechanistic insights that underlies phenotypes observed in myeloid cancer patient. The mechanisms uncovered by this study will provide valuable

knowledge of how early hematopoietic stem cell fate is committed and inform therapeutic intervention strategies through cellular reprogramming and cell fate intervention.

CHAPTER 2: MATERIALS AND METHODS

2.1 Materials and methods for chapter 3

Cells and reagents

Human mobilized peripheral blood CD34+ HSPCs were purchased from the Fred Hutchinson Co-operative Center for Excellence in Hematology (Seattle, WA, USA), and were obtained from multiple healthy donors. CD34+ HSPCs were expanded in StemSpan SFEMII base media supplemented with CC110 culture supplement for 3-5 days prior to screen (Stemcell Technologies, Vancouver, Canada). Human cytokines SCF, IL3, IL6, and EPO were purchased from Peprotech Inc (Princeton, NJ, USA). Oxyphenonium bromide was purchased from Sigma (St. Louis, MO, USA). The following antibodies were used for western blots: anti-PTEN (1:1000, Cell Signaling #9552, Danvers, MA), anti-CUX1 (1:1000, Santa Cruz #sc514008, Dallas, Texas), anti-GATA1 (1:1000, Abcam #ab181544, Cambridge, UK), anti-beta-actin (1:10000, Santa Cruz #sc47778, Dallas, Texas).

gRNAs, Cas9, and Neon transfection

Guide RNAs and Cas9 were purchased from the Synthego corporation and gRNAs designed using their bioinformatics tools. A single gRNA per gene was employed. Electroporation transfection was performed on the Neon Transfection System (Thermofisher Scientific, Waltham, MA, USA) as previously reported.[209] Briefly, electroporations were performed at settings of 1600 volts, 10 ms pulse length, 3 pulses. Prior to transfection, 0.71 ul Cas9 (20uM) was mixed with 2.39ul gRNA (30uM) and 0.9 ul Buffer T, and RNPs allowed to form for at least 15 minutes at RT. 200 000 CD34+ cells in 8ul of Buffer T were then added

to the RNPs, and 10 ul of the mixture electroporated and immediately cultured in SFEMII +CC110 to recover for 24 hours before use in proliferation and erythroid differentiation assays. AAVS1 gRNA was used as a negative control.[210] gPTEN was included as a positive control for increased proliferation, and gGATA1 was used as a control for decreased differentiation.[211,212] A complete list of gRNAs and primers used for Sanger sequencing can be found in the supplemental data file.

Proliferation assay

Proliferation cultures were seeded with 10 000 cells per well in 96-well plates, containing 200ul of SFEMII +CC110 plus 10 nM IL3 and 10 nM IL6. Proliferation assays were performed at days 3, 5, and 7 using the CellTiterGlo 2.0 Cell Viability Assay (Promega, Madison, WI, USA) in duplicate. Assay was performed as per manufacturer's protocol, with 25 ul of cultured cells and using freshly diluted ATP solutions (Sigma) for standard controls. Each group of gRNAs was repeated so that all target genes were tested in 3-4 separate biological replicates, with 4 assay replicates for each target gene in each biological replicate (total of 12-16 data points for each target gene).

Erythroid differentiation assay

Erythroid differentiation cultures were seeded with 25 000 cells per well in 96-well plates in SFEMII base media plus 25 nM SCF, 10 nM IL3, 10 nM IL6, and 6 units/mL EPO. Cultures were then grown for 14 days, expanding into 24-well plates and splitting as necessary to avoid confluence. At 14 days, cells were subjected to flow cytometry for erythroid markers

CD71-BUV395 (BDBiosciences, cat# 743308, San Jose, CA, USA), GlyA-FITC (BioLegend, cat# 349104, San Diego, CA, USA), and LIVE/DEAD Fixable Near-IR Dead Cell Stain (Thermofisher Scientific). All flow cytometry performed on an LSRII instrument (BD). Each group of gRNAs was permutated same as the proliferation assay but without assay replicates, resulting in n = 3-4 for each target gene.

Combined proliferation and erythroid differentiation scores

For erythrocyte differentiation and proliferation scores, I normalized the experimental results of each gene in both assays in each well by subtracting from AAVS1 control. Erythroid signs were inverted so that a higher score is associated with increased proliferation and decreased erythroid differentiation. Then I obtained an average erythrocyte differentiation and proliferation score by taking the mean of each gene's results across all replicate plates. I then combined the average erythrocyte differentiation and proliferation scores into one table and performed min-max normalization onto a 0-1 range, in order to remove the effect of directionality and unify the two scores onto the same scale. Finally, I summed the normalized proliferation and erythrocyte differentiation score to obtain the "combined experimental score", which is a unified measure of the likeliness for each gene to be a tumor suppressor. To test whether the proliferation and erythrocyte differentiation results for each gene are significantly different from those of the AAVS1 control, I performed a non-parametric Mann-Whitney test on the results of each gene across all replicate wells compared to AAVS1. The p values were multiple hypothesis corrected using Storey's q value (FDR < 0.12).

Machine Learning Classifier

I used a classification random forest model to predict putative tumor suppressors on chromosome 7, based on publicly available genome-wide screening data in human cell lines. For our training data, we compiled a list of genome-wide screening data, Subsequently, I applied the following filters to pre-process our training data: 1) predominantly retained the screens related to hematological malignancies including AML, CML and Burkitt's Lymphoma; and 2) filtered out two screens with too many 0 values (90%) and missing values (66%). This led to screening data from 8 publications spanning 24 cell lines. Furthermore, I added mutational signature data from the Davoli et al. 2013 study which were shown to have the best performance in predicting tumor suppressors using a LASSO regression model.[42] I then used a k-nearest-neighbor algorithm to impute the remaining missing values. For the "ground truth" column used in training our model, we used annotation from Cancer Gene Census (CGC),[213] which labelled 315 canonical tumor suppressors genome-wide. I then split the training and testing data. Our testing data consisted of all the protein-coding genes on chromosome 7, and the training data are all the protein-coding genes on all other chromosomes. For the training data, I did 100 iterations of bootstrapping in order to remove the effect of randomness and achieve training-testing data balance. In each bootstrap, I randomly sampled genes labelled as non-tumor suppressors from CGC and matched them with a comparable number of genes labelled as tumor suppressors. For each bootstrap, I performed hyperparameter tuning of the random forest model and selected the combination of hyperparameters that gave the smallest out-of-bag (OOB) error rate. I then performed prediction on the testing data using the 100 tuned models and obtained a binary result for each gene (1 for putative tumor suppressor and 0 for non-tumor suppressor). I ranked the

chromosome 7 genes based on the frequency of being labelled as a tumor suppressor in the 100 bootstraps to obtain the final list. Performance evaluation was performed on training data. Across 100 bootstraps, we obtained an average AUC of 0.777(0.747 – 0.806, 95% CI). We achieved an average of 71.8% accuracy, 73.4% precision, 59.5% sensitivity, and 82.0% specificity.

2.2 Materials and methods for chapter 4

Co-immunoprecipitation

100 × 106 K562 cells were spun down for a CUX1 pulldown and a control IgG pulldown each. Cells were lysed in hypotonic buffer (5 mM EDTA, 5 mM EGTA, 5 mM Tris–Cl) with protease inhibitor added (Roche complete mini-EDTA free 11836170001). Pellets were passed through a 20-gauge needle 10 times, incubated on ice for 10 minutes and spun down at 600 g for 8 minutes at 4°C. The supernatant was removed, and the pellet was resuspended in RIPA buffer (Boston BioProducts BP115) with protease inhibitor added (Roche Complete 5892953001). Protein lysates were again passed through a 27-gauge needle, incubated on ice and subsequently spun down at 14000 rpm for 15 minutes at 4°C. The supernatant was collected, and RIPA buffer was added to a final volume of 30 mL. 12 µg of CUX1 antibody (B-10 Santa Cruz sc-514008) and mouse IgG (Santa Cruz sc-2025) antibody were added to the lysate and incubated overnight on a rocker at 4°C. 150 uL Protein A/G Plus agarose beads (Santa Cruz sc-2003) were added the next day and incubated at 4°C on a rocker for 1 hour. The immunoprecipitated proteins were washed twice with cold RIPA buffer followed by a final wash with cold PBS. Proteins were eluted by resuspending the beads in 2X loading buffer and sent for mass spec analysis.

Sample preparation for LC–MS/MS

Co-immunoprecipitate samples were brought to 1X and 40 uL was loaded onto 12% MOPS buffered 1D SDS-PAGE gel (Invitrogen NP0341BOX) and run at ~ 200 V for ~ 10 min, resulting in a ~ 2 cm gel plug. The gel was stained with Imperial Stain (Thermo Fisher #24615) for 1 hour at room temperature. Gel plug trypsin digestion was adapted from methods previously published.79,80 Digested peptides were cleaned up via C18 spin columns (Thermo Fisher #89870).

LC–MS/MS via MaxQuant

LC–MS/MS was performed using adapted methods previously published79. Electrospray tandem mass spectrometry (LC–MS/MS) was performed at the Mayo Clinic Proteomics Core on a Thermo Q-Exactive Orbitrap mass spectrometer, using a 70,000 RP (70 K Resolving Power at 400 Da) survey scan in profile mode, m/z 340–1800 Da, with lockmasses, followed by 20 MS/MS HCD fragmentation scans at 17,500 resolutions on doubly and triply charged precursors. Single charged ions were excluded, and ions selected for MS/MS were placed on an exclusion list for 60 seconds.

Mass spectrometry database searching and analysis

Tandem mass spectra MS/MS samples were analyzed using MaxQuant (version 1.6.17.0). MaxQuant was set up to search the 211102_Uniprot_Human_5640.fasta database assuming

the digestion enzyme strict trypsin. MaxQuant was searched with a fragment ion mass tolerance, and a parent ion tolerance of 20 PPM. MQ 1FDR results file (proteingroups.txt) was processed in Perseus (version 1.6.14.0). Proteins were filtered out which included "identified by site", "reversed", and "potential contaminants", log2 transformed, imputed via default settings, and annotated against the human database. P-values were determined by Student's t-test within Perseus and a significance cutoff was applied if CUX11/IgG ratios were above NegLog10 P-value >= 1.3 and fold-change above 20% or log2 >= 0.26. Proteins only detected in CUX1 immunoprecipitates were also determined significant.

Cell culture

K562 cell lines were obtained from Dr. Michelle Le Beau's lab (University of Chicago) and were authenticated by STR analysis (ATCC). Primary human CD34+ peripheral blood mononuclear-stem cells were obtained from the Fred Hutch Hematopoietic Cell Procurement and Resource Development Center (Seattle, WA). K562 cells were grown in RPMI 1640 media (Gibco 61870127) supplemented with 10% FBS and 1X Antibiotic-Antimycotic (Gibco 15240062). Primary CD34+ cells were grown in StemSpan SFEMII media (STEMCELL Technologies 09655) supplemented with 1X StemSpan CC110 cytokine cocktail (Stemcell Technologies 02697).

Ribonucleoprotein (RNP) transfection

gHPRT and gCUX1 K562 cell lines were described previously.[81] For primary CD34+ HSPCs, cells were transfected with ribonucleoprotein complexes carrying the same gRNA

sequences as used in K562 for exon 4 of CUX1 (5'- UGCACUGAGUAAAAGAAGCA-3')[214] or intron 2 of HPRT (5'-GCAUUUCUCAGUCCUAAACA-3') (Synthego) using the Neon transfection device (Thermo Fisher) with the following parameters: 1600V, 10ms, 3 pulses.[215] Editing efficiency was determined 72 hours post-transfection using TIDE (https://tide.nki.nl/).[216] The editing efficiencies for the transfected cell population replicates used for experiments are: gCUX1 replicate one 47%, replicate two 52%; gHPRT replicate one 79%, replicate two 72%.

ChIP-seq library preparation and sequencing

Chromatin was fixed from 100x106 gHPRT and gCUX1 transfected K562 cells using 1% formaldehyde for 10 minutes at room temperature and stopped by the addition of 0.125 M glycine. For SMARCA4 ChIP, protein cross linking was performed first. Cells were washed 3 times with 1X PBS at room temperature. 10 mL of PBS/MgCl2 were added to the cells after final PBS wash. 80 uL of 0.25M DSG-disuccinimidyl glutarate (Thermo Fischer 20593) was added and incubated for 45 minutes at room temperature. Cells were washed 3 times with 1X PBS and followed by DNA crosslinking with 1% formaldehyde as described above. Fixed chromatin was then sonicated (Bioruptor) for 10 minutes in 30 seconds on/off pulses two times for a total of 20 minutes, with vortexing in between. CUX1-specific antibodies were generated, characterized and validated as described by Imgruet et al. 2021.[81] Immunoprecipitation was performed using dynabead protein G magnetic beads (Thermo Fischer) and 6 ug of anti-CUX1 (PUC, Poconos)/20E6 cells, 5 ug/20E6 anti-SMARCA4 (Abcam ab110641), anti-H3K27ac (Abcam, ab4729) or anti-H3K4Me1 (Abcam, ab8895). Following elution, samples were treated with RNase A and proteinase K before crosslink

reversal. DNA was purified using a PCR purification kit (Qiagen). Libraries were prepped using the Ovation Ultralow Library Kit Tecan Genomics Inc (0344NB-32) and size selected using SPRIselect beads (Beckman Coulter B23317). Illumina HiSeq was used to perform 50 bp single-end sequencing on the libraries. Two biological replicates were performed for each sample.

CUT&RUN library preparation and sequencing

CUT&RUN was performed as described by Skene and Henikoff 2017.[217] using the direct ligation method for mammalian cells. Briefly, 5x105 cells were harvested from CD34+ HSPCs expanded for 48 hours post-thawing and bound to ConA-coating beads by rotation for 10 minutes at room temperature. Cells were permeabilized (20 mM HEPES pH 7.5, 150 mM NaCL, 0.5 mM Sperimidine, Roche Complete EDTA free 5892953001, 0.05% w/v digitonin) and incubated overnight at 4°C with anti-CUX1 (PUC) 1:50, anti-SMARCA4 (Cell Signaling, 49360s) 1:100, or anti-GATA1 (Abcam, ab181544) 1:100 antibodies. Protein A/G-MNase beads were added and placed on a tube rotator for 1 hour at 4°C. MNAse bound DNA was cleaved and released by adding 1X pA-MNAse mix containing CaCl2 at 0°C for 30 minutes, STOP buffer was added and CUT&RUN fragments were released by incubating for 30 minutes at 37°C. Library end repair, ligation, and amplification were performed using the Ovation Ultralow System V2 kit (Tecan Genomics Inc. 0344NB-32) and amplified by PCR with the following parameters: 1 cycle of 72°C 2 minutes, 95°C 3 minutes, followed by 13 cycles of 98°C 20 seconds, 65°C 30 seconds, 72°C 30 seconds, and a final extension at 72°C for 1 min. Libraries were cleaned up using MinElute PCR purification kit (Qiagen) and

a left-sided size selection using SPRI beads (Beckman Coulter B23317). Final libraries were analyzed by Bioanalyzer (Agilent) prior to sequencing.

ATAC-seq sample preparation and sequencing

ATAC-sequencing was performed according to a published protocol.[218] For all experiments, K562 cells were harvested from cultures at ~60% confluency and primary CD34+ HSPCs were harvested 48 hours post transfection. For both K562 and primary CD34+ HSPCs, 50,000 cells were lysed using the following buffer: 10 mM Tris-HCL, pH 7.4, 10 mM NaCL, 3 mM MgCl2, and 0.1% IGEPAL CA-630. Cells were transposed using a 1X concentration of Nextera Tn5 Transposase (Illumina) for 30 minutes at 37 °C with shaking at 500 rpm. Following transposition, DNA was purified using the MinElute PCR Purification Kit (Qiagen). DNA was amplified for 5 initial cycles using the custom Nextera barcoded PCR primers with the following parameters: 1 cycle of 72 °C for 5 minutes and 98 °C for 30 seconds, followed by 5 cycles of 98 °C for 10 seconds, 63 °C for 30 seconds, and 72 °C for 1 min. Following the initial 5 cycles of PCR, the additional number of cycles needed was determined by qPCR as previously described.[218] Following the additional PCR cycles, DNA was obtained using the MinElute PCR Purification Kit and analyzed by Bioanalyzer (Agilent) prior to sequencing.

ChIP-seq and CUT&RUN analysis

For the ChIP-seq analysis using K562 cells, sequenced samples were trimmed using Cutadapt (version 4.2.0).[219] We aligned single-end reads to hg19 using bwa (version 0.7.17) and called peaks using MACS2 (version 2.1.0) with input control.[220,221] All peak calling was performed according to the ENCODE standards using an irreproducible discovery rate (IDR) of 0.05.[222] Non-uniquely mapped reads and reads mapped to ENCODE blacklist region[223] composed of artificially high regions of the genome were discarded. Coverage files were generated using deepTools (version 3.5.1) and visualized using IGV (version 2.8.10).[224,225] ChIP-seq for CUX1, SMARCA4 and histone marks H3K27ac were performed at the McNerney lab. GATA1 (ENCSR000EWM) and RUNX1 (ENCSR414TYY) ChIP-seq data were obtained from ENCODE. We assigned peaks to the single nearest transcription start site (TSS) within 1 Mb using GREAT (version 4.0.4).[226] Bed files were analyzed using Bedtools (version 2.30.0).[227] Significance of overlap of binding sites between two ChIP-seq experiments was calculated using the hypergeometric test with makeVennDiagram() from ChIPpeakAnno package (version 3.32.0), with options: " totalTest=totalTest,scaled=FALSE, euler.d=FALSE,method = "hyperG".[228] We used MEME-ChIP for motif discovery using the classical mode.[229,230] Summits of CUX1 and SMARCA4 binding sites were calculated and extended in both direction by 250 base pairs as the sequence input. Accessible chromatin sites obtained from the K562 gHPRT ATAC-seq were used as the background model to increase the statistical power pf motif discovery. Differential motif analysis was performed using AME.[231]

For CUT&RUN analysis for CUX1, GATA1 and SMARCA4 in primary human CD34+ HSPCs, all analysis methods and parameters are the same as in ChIP-seq (based on how other people analyze CUT&RUN in the literature), except the sequencing reads are paired end.

ATAC-seq analysis

For both K562 and human CD34+ HSPC ATAC-seq analysis, sequenced samples were trimmed using Cutadapt (version 4.2.0).[219] We aligned paired-end reads to the human hg19 genome using bwa (version 0.7.17) and called peaks using MACS2(version 2.1.0) with " --nomodel, -- shift -75, and -- extsize 150 " options.[220,221] Non-uniquely mapped reads, mitochondrial reads, and reads mapped to the ENCODE blacklist region[223] were discarded. Coverage files were generated using deepTools (version 3.5.1) and visualized using IGV(version 2.8.10).[224,225] Differentially accessible regions in gCUX1 vs. gHPRT samples were identified using csaw using a 2-fold enrichment threshold and FDR smaller than 0.05.[232] We chose csaw because it is an unbiased approach that scan through the whole genome using a sliding window approach, rather than depending on pre-called peaks. The unbiased csaw approach is better at picking up more subtle differential changes that do not reside in the pre-called peak regions. Bed files were analyzed using Bedtools (version 2.30.0).[227] For integration with RNA-seq, ATAC-seq peaks are identified to be the significant ATAC peaks called by csaw[232] within 1 Mb window from the TSS of the differentially expressed genes(FDR<0.1, |Log2FC|>0.75) identified from RNA-seq in shCUX1 vs shControl.[79] 406/432 DEGs have significant ATAC peaks within 1 Mb window from their TSS and are thus retained for this analysis. The ATAC peak with highest Log2FC for each gene was selected.

Analysis of chromatin accessibility at cell type specific enhancers

We downloaded the cis-regulatory element annotation map generated by Zhang and Hardison, 2017[233] for primary human hematopoietic cell types including HSC, MEP, GMP, CLP, erythrocyte, megakaryocytes, neutrophils, monocytes, B cell, NK cells, CD4+ and CD8+ T cells from the Validated Systematic IntegratiON of hematopoietic epigenomes (VISION) data portal. (https://usevision.org/). We retained all the genomic intervals identified as enhancers for each cell type, including E: enhancer like; EN: enhancer like, nuclease accessible; EN_A: enhancer like, nuclease accessible, active; E_A: active enhancers; BE: bivalent enhancers; CNE_T: CTCF bound, nuclease accessible, transcribed enhancers; TE_A: transcribed active enhancers; TE: transcribed enhancers. We eliminated all the enhancer elements that are annotated ambiguously as promoter-like. For each progenitor cell type, we eliminated the enhancer elements that are shared in HSCs in order to obtain a list of enhancers that are unique in each specific progenitor cell type. Then we calculated the normalized chromatin accessibility from our CD34+ HSPC ATAC-seq data gHPRT and gCUX1 at the cell type specific enhancers. For plotting, the cell types are merged into lineages: Erythroid (MEP + megakaryocytes + erythrocytes), Myeloid (GMP + neutrophils + monocytes) and Lymphoid (B cells + NK cells + CD4+ and CD8+ T cells). The negative control is a list of 10,000 randomly sampled enhancers that did not appear in any of the cell type specific enhancer lists.

Annotation of peaks with chromatin state

ChIP-seq, CUT&RUN and ATAC-seq peaks were annotated with chromatin state using publicly available data. K562 chromatin state prediction was obtained from UCSC genome

browser chromHMM track, which uses hidden Markov model analysis of eight chromatin marks and CTCF ChIP-seq data.[234,235] Primary human CD34+ HSPC chromatin state data was obtained from NIH Roadmap Epigenomics Mapping Consortium (EP50 primary hematopoietic stem cells G-CSF-mobilized female chromHMM track).[236] The database also used hidden Markov model analysis of six chromatin marks and DNase I hypersensitivity data. To establish the chromatin state of genomic sites, we used Bedtools intersect (version 2.29.0) to obtain the overlap of each ChIP site with chromHMM annotations.[227]

Hi-C analysis

Hi-C data from CD34+CD38- primary human HSPC was obtained from a published study.[237] We intersected CUX1-bound promoters (defined as CUX1 binding sites in human CD34+ primary HSPC CUT&RUN that fall within 2 kb from the TSS) with the 2,684 chromatin loops called by Zhang et al 2020.[237] 272 loops were found to contain CUX1-bound promoters. We then found the interacting regions of these 272 loops and defined them as the regions that contain putative enhancers in contact with CUX1-bound promoters. Normalized ATAC seq reads (RPKM) in gHPRT and gCUX1 samples on these regions were calculated using deepTools (version 3.5.1).[224] As the negative control, we size-matched and randomly sampled 272 regions that are not in contact with any CUX1-bound promoters.

Murine HSPC fate prediction

We obtain the scRNA lineage tracing data from Weinreb et al. 2020,[238] where murine Lin-Sca^{high} Kit+ HSCs were clonally traced by expressed DNA barcodes so that the terminally differentiated daughter cell fates are linked with ancestor HSC single cell transcriptomes. We downloaded the in vivo normalized count matrix and metadata containing the single cell clonal identities from the GEO database (GSE140802). Seurat V4 was used to import, preprocess and analyze the data.[239] Ancestor HSCs and daughter cells were assigned to their clonal identities. We filtered the cells that do not belong to any clones and the HSCs that do not have any daughter cells. All terminal cell fate annotations were stored in a list. We then looped through this list and determined the most common (if there is one) cell fate. For example, a clonal lineage with the cell fates A, A, B will be determined as being a clonal lineage A, while one with the cell fates A, B will be listed as ambiguous. This gave us 1,523 cells with unique terminal identities after removing cells with ambiguous or undifferentiated cell fates. The remaining cells contained basophil, dendritic cells, monocytes, neutrophiles, B cells, and erythrocyte progenitors. After building datasets with the gene expression levels matrix on one side and the cell fate on the other, we ran different Python scikit-learn machine learning models and graded their accuracy (We chose F1 score as the measurement for prediction performance due to label inbalance and the better control on type I and II errors) to determine how informative different sets of genes were in determining cell fate.[240] The two models we used were "LogisticRegression" and deep neural network (implemented by MLPClassifier). For the MLPClassifier, hyper parameter tunning using "GridSearchCV" was performed to identify the best parameters on each dataset. We then ran the prediction model and compared the cell fate prediction accuracy for different gene sets. Since some of these gene sets came from experiments on human and our training data is from mouse, we had to

convert the gene sets from human to the corresponding mouse gene names using the R package biomaRt.[241] For each gene set, to reduce sampling bias, we performed 50 bootstrap analyses and took the average and standard deviation of the scores. The average accuracy for each of our models was recorded. From previously published studies, we obtained genes corresponding to PU.1 (n = 2,074) and RUNX1 binding sites (n = 5391), and PU.1 and RUNX1 bound genes that are differentially expressed after they were lost in HSPCs (n = 336 and n = 325 respectively).[242–245] We performed two-sided Wilcoxon rank-sum test on the F1 scores between our experimental datasets and the most variable genes, randomly selected genes, and mouse transcription factors obtained from AnimalTFDB 3.0.[246]

Creating customized reference genome that differentiates the *CUX1* and *CASP* transcripts

From NCBI RefSeq database (https://www.ncbi.nlm.nih.gov/refseq/), I obtained the transcript identifier of the three *Cux1* transcripts (NM_001291233.1, NM_009986.4, NM_001291234.1), and the four *Casp* transcripts (NM_001291239.1, NM_001291240.1, NM_198602.3, NM_001291238.1). I then downloaded the fasta and gtf annotation file for mouse genome version mm10 from Ensemble database (https://useast.ensembl.org/info/data/ftp/index.html). All *Cux1* and *Casp* transcripts are named *Cux1* in the gene symbol column in the generic gtf file. Therefore, I changed the gene symbol form *Cux1* to *Casp* for the four *Casp*-specific transcripts. The, I used STAR[247] built-in reference genome generator to generate the customized genome which differentiated the *Cux1* vs *Casp* transcripts. The parameters are as follows: *STAR --runThreadN 1 --runMode*

genomeGenerate --genomeFastaFiles mm10.fa --sjdbGTFfile mm10_custom.gtf. The same procedures are done to create the equivalent human hg19 reference genome.

Single cell RNA-seq library preparation and sequencing

Two constructs of doxycycline-inducible shRNA[79] targeting different Cux1 exons and one control construct of shRNA targeting renilla were transfected separately to bone marrows of healthy donor mice. $Cux1^{mid}$ targets exon 5 which is shared by all *Cux1* and *Casp* isoforms. The residual CUX1 protein level is 54% +/-17%. $Cux1^{low}$ targets the 3' untranslated region of exon 24, which is only shared by *Cux1* isoforms and not *Casp*. The residual CUX1 protein level is 12% +/-9%.[79] The transfected bone marrows were transplanted to bone marrow irradiated mice. Three littermate mice were transfected with each shRNA construct. Then, these recipient mice were allowed to recover for four weeks. After the recovery period, they were fed doxycycline diet for five days in order to induce the shRNA expression. Initially, mice are euthanized to harvest their legs, hips, and arms. The bone marrow is then isolated, crushed, and lysed, with the resulting cell suspensions from three mice in each genotype condition combined. Cell counts are performed using a cellometer, with the following counts obtained: Renilla control 1.57^7 cells with 86.7% viability, $Cux1^{mid}$ 1.53^7 cells with 83.8% viability, and $Cux1^{low}$ 1.16^7 cells with 82% viability. Next, we performed the lineage depletion process. A fraction of the undepleted cells is set aside for flow cytometry confirmation of lineage depletion. The samples are then incubated in a 4°C fridge for 10 minutes. Afterward, a 3mL buffer containing EDTA is run through columns, followed by a 1mL buffer addition to each sample, which is then applied to the columns. The flow-through, containing lineage-depleted cells, is collected for further analysis and to obtain GFP+ cells.

The lineage-depleted cells are centrifuged, resuspended in 1mL, and recounted on the cellometer, yielding the following counts: Renilla control $1.89x10^7$ cells with 91.1% viability, $Cux1^{mid}$ 1.64^7 cells with 91.2% viability, and $Cux1^{low}$ 5.71^6 cells with 93.6% viability. Subsequently, cells are resuspended in Fc block and incubated for 10 minutes on ice, shielded from light. Cells are then stained with a lineage cocktail and incubated for 30 minutes at room temperature in the dark. This is followed by staining with a master mix, which involves spinning down the cells, resuspending in the designated volume of staining mix, and incubating again for 30 minutes at room temperature, protected from light. Cells are then resuspended in 400 µL of buffer (the volume may vary) and stored on ice. On the day of the final procedure, cells are spun down 30 minutes prior to the sorting time. Cells are then resuspended in a 1:10,000 dilution of live/dead stain from the -20°C storage and incubated for 15 minutes at room temperature. After a final spin down, cells are resuspended in 400µL of buffer. The final step is the sorting process, during which 2 million kit+, GFP+, lineage-, live cells are collected for subsequent experiments. Only hematopoietic stem cells and early progenitors were retained using this sorting strategy. The three cell populations (shRenilla control, $shCux1^{mid}$ and $shCux1^{low}$) were sequenced individually using 10X genomics single cell RNA sequencing technology.

Single cell RNA-seq analysis

After sequencing, UMI counts were obtained for gene expression via gene-barcode matrix with 10x genomics' Cell Ranger (version 7.0, Chromium Single Cell V(D)J Reagent Kits with Feature Barcoding technology for Cell Surface Protein, Document Number CG000186 Rev A, 10x Genomics, (2019, July 25). a set of analysis pipelines that process Chromium single-cell RNA-seq output. I used Seurat (Version 4) for the ensuing data preprocessing.[239]

To get rid of non-expressed genes, I kept all genes expressed (defined by nonzero counts) in ≥10 cells. To retain intact and healthy singlet cells, I applied the following filter: 1) At least 200 genes expressed are required, indicating a intact cell. 2) Cells must express < 8% mitochondrial reads, indicating non-bursting cells 3) library size of cells within 3 standard deviation around mean value. This effectively removed droplet contains small fragments and congregated cells. Global-scaling normalization method "SCTransform" was performed on the filtered data at the next step. I performed cell cycle regression in the next step. Because HSCs is quiescent versus more proliferative progenitors downstream, early HSCs tend to reside in G1/G0 phases while more proliferative cells tend to reside in the cycling G2M or S phase. I don't want to lose information that could differentiate stem vs non-stem cells. Therefore, I am just regressing out G2M vs S phase. Signals separating non-cycling cells and cycling cells will be maintained, but differences in cell cycle phase amongst proliferating cells (which are often uninteresting), will be regressed out of the data. Dimension reduction was performed using both PCA and UMAP approaches. Unsupervised clustering was then performed using KNN algorithm, multiple resolutions of KNN were performed, and the resolution that returns stable number of clusters was determined using clustree R package.[248] Cluster identity was annotated by manually annotating the top differentially expressed up-regulated genes in each cluster vs. all the other clusters. To increase the robustness of the cell type annotation, an automated approach using publicly available bulk RNA-seq and microarray data in different hematopoietic cell types was also performed using singleR R package.[249] Results from both the manual and automatic approaches were cross checked and are mostly consistent. Droplets containing doublet cells were removed using DoubletFinder R package, which were shown to be the top packages in terms of accuracy in finding doublets.[250] Afterwards, the Renilla control, Cux^{mid}, and $Cux1^{low}$ samples were integrated using Seurat (V4.0),[239] in order to remove bias due to sequencing depth and enable

comparison across samples. Pseudotime inference and trajectory analysis was performed using Slingshot and TradeSeq.[251,252] because of their benchmarked high performance and suitability for multi-trajectory differentiation, which is the topology of a typical hematopoietic differentiation dataset.[253] Differential abundance test comparing the sub-cluster cell numbers across different conditions is calculated using EdgeR framework.[254] Three pseudo-replicates were created for each condition(Ren, Mid, Low) by random sampling. Each cluster is equivalent to a gene and cell number of each cluster is equivalent to the raw reads aligned to each gene for a typical RNA seq differential expression analysis. Clusters with low cell numbers (< 50) are excluded, equivalent to getting rid of lowly expressed genes. Dispersion was estimated between samples and replicates. Test for differential abundance was implemented using the negative binomial model of EdgeR. Composite effect was controlled by eliminating the most abundant cluster and repeat the test to ensure the results still hold true. All other downstream analysis are performed using Seurat and self-generated code, which is available on Github (https://github.com/liuweihanty/single_cell_RNA_Cux1)

2.3 Chapter 5 Materials and methods

GATA1 CUT&RUN library preparation and sequencing for K562

The protocol is the same as the CUT&RUN protocol for CD34+ HSPC outlined previously in this section. We also collected DNA from 5 x 10^5 K562 cells, and used anti-GATA1 (Abcam, ab181544) 1:100 antibody. All computational analysis is also the same as previously outlined.

CHAPTER 3: SYSTEMIC DATA MINING OF GENOME-SCALE SCREENING DATA IDENTIFIED PUTATIVE TUMOR SUPPRESSORS ON CHROMOSOME 7

3.1 INTRODUCTION

-7/del(7q) is one of the most frequent and serious cytogenetic abnormalities in myeloid malignancies, including MDS and AML. -7/del(7q) is associated with worse prognosis and chemotherapy resistance, regardless of age and disease.[255] Mechanistically, -7/del(7q) contributes to the initiation of transformation, and is detected in clonal hematopoiesis.[31,32] Despite the clinical implication of -7/del(7q), the specific mechanisms of how they contribute to malignant transformation are not understood. It has been suspected that key TSGs encoded on chromosome 7 commonly deleted regions (CDR) might contribute to transformation, but identifying these genes has been challenging due to technical and biological limitations. The classical approach to look for recessive TSGs from CDRs have been ineffective and only limited number of TSGs have been found on chromosome 7, including *CUX1* through mapping CDR regions,[49] *EZH2* through searching for second hit mutations[44] and *SAMD9/SAMD9L* via microarray hybridization.[256] A systemic approach to identify putative tumor suppressors on chromosome 7 CDRs is needed.

The advance in genome-wide perturbation screens in the last decade provided us with rich data source to computationally mine the TSG phenotypes and predict putative TSGs. These methods use genetic perturbation method to systemically knock out/in or up/down tune the expression level of hundreds to thousands of genes in a high throughput fashion by plate based or pooled methods using DNA barcodes. The targeted cells could grow underneath a biological treatment of interest, such as drug treatment, cell competition, or simply a survival essay without external stimuli. Subsequently, the biological effect of the challenge is measured by various assays. Early high throughput genetic screen methods include gene

trapping mutagenesis, where target gene expression is randomly disrupted by a transgene vector with a reporter that can simultaneously tag the identity of the target gene.[257] Random mutagenesis induced by chemicals are also used for high throughput screens. However, these methods are painstakingly time-consuming and difficult to implement experimentally. The advent of RNAi method significantly simplified the way to knock down gene expression and led to a wave of both *in vitro* and *in vivo* studies identifying novel functions of genes in a high throughput fashion.[258] Subsequently, CRISPR/Cas technology has emerged as the go-to tool for large scale screen due to its ease of use and flexibility. CRISPR toolbox allows researchers to knockout target genes with Cas9 endonuclease cutting,[259] or by attached an effector protein to deactivated Cas (dCas) protein to tune the expression of the target gene either up (CRISPRa) or down (CRISPRi), without inducing DNA double stranded breaks.[260,261] With such diverse and flexible genome engineering toolbox available, there have been numerous CRISPR genome-wide perturbation studies for tumor suppressor function. The idea is to transfect a pooled population of cells with sequence-specific gRNAs that target thousands of genes. Through careful titration, each cell receives one copy of the gRNA and thus perturbation on one gene. The perturbation induced effect is measured by sequence-based counting of gRNAs at the end of assaying period.[262] The readout is reflected in CRISPR score (CS), which measured the log transformed sgRNA abundance for each guide in the start and end of the screen. Genome-wide loss-of-function screens by either knocking down or inhibiting TSG expression in cancer cell lines led to cell overgrowth, while knocking in or upregulating TSG generally leads to decreased cellular proliferation (**Figure 3.1**).

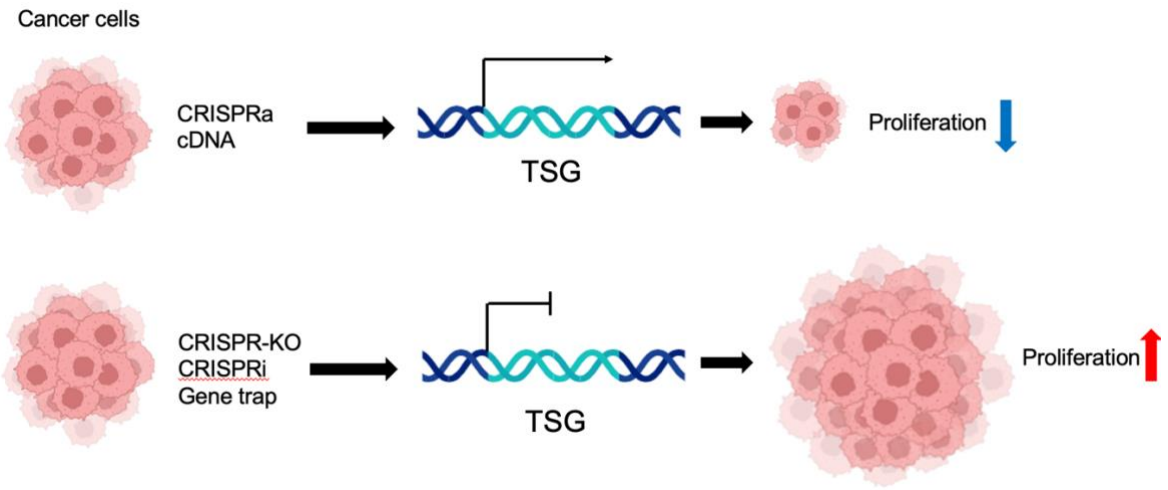


Figure 3.1: Schematic of the effect of different types of genome-wide perturbation screens on cancer cell growth. Top panel: upregulating gene expression by CRISPR-activation (CRISPRa) or cDNA library overexpression. Bottom panel: downregulating gene expression using CRISPR knockout (KO), CRISPR inhibition (CRISPRi) or gene trap mutagenesis screens. The graph was created using biorender (https://www.biorender.com/).

The bountiful genome-wide perturbation data sets provide a rich resource to apply supervised machine learning and learn the pattern on the effect of perturbing the TSGs. Random forest is a popular supervised machine learning method utilizing ensemble trees. The ability of random forest to handle high-dimensional data, capture complex interactions among genes or features, and provide feature importance rankings makes it a valuable tool in deciphering the intricacies of genetic data.[58] In this project, I collected a wide-range of genome-wide perturbation data in hematopoietic cancer cell lines and applied the random forest model to learn how canonical TSGs behave in these screens, and subsequently used the trained model to predict TSG activities for all chromosome 7 genes. Before the commencement of this project, we were not aware of other studies that leveraged supervised machine learning on these screen data to predict TSG activities.

3.2 RESULTS

3.2.1 Compiling genome-wide screening data suitable for identifying TSGs.

To select the proper training data for our machine learning classifier, I focused on genome-wide screening data in human cancer cell lines. For the training data, I compiled a list of data composed of gene trap mutagenesis, cDNA library overexpression, CRISPR-KO, CRISPRi and CRISPRa screens. The training data contains proliferation scores for each gene screened, which measures the relative abundance of the cells edited on each gene at the start and end of each screen. For example, for CRISPR screens, this is the normalized ratio of gRNA abundance at the end and start of the screen. Subsequently, we applied the following filter to preprocess our training data:

1) Only retain the screens related to hematological malignancies, including acute AML, CML and Burkitt's Lymphoma. We reason that this will improve the accuracy of categorizing tumor suppressors, as the same gene could function differently in different types of cancers.

2) Filtered out two screens with too many 0 values (% of zeroes must be smaller than 90%) and missing values (missing values must be smaller than 66%). Too many zeros could indicate bad quality of the screen experiment like poor editing efficiency, and too many missing values will make the information extracted from the screen unreliable. The screens left all have less than 60% zeroes and less than 66% missing values.

This led to screening data from eight publications spanning 24 different types of cell lines, summarized in **Table 1**. Furthermore, we added mutational signature data from the Davoli et al. 2013 study which were shown to have the good performance in predicting tumor suppressors using a LASSO regression model.[42] In this study, the authors used pan-cancer

mutational dataset and found that these parameters have the best performance to predict TSG vs neutral genes and oncogenes:1). ratio of loss-of-function (LOF) vs benign mutations. 2) Splicing vs benign mutations 3) High functional impact (HiFI) missense vs benign mutations 4) deletion frequencies.[42] I incorporated the score of these mutations at each genomic loci in my classifier.

Table 2. Genome wide screening data[263–271] used in the machine learning classifier

Source	Cell Line	Cancer Type	Data Type
Blomen et. al. 2015	HAP1	AML	Gene Trap Score
Blomen et. al. 2015	KBM7	AML	Gene Trap Score
Gilbert et. al. 2014	K562	CML	CRISPRa Score
Wang et. al. 2015	KBM7	AML	CRSIPR-KO Score
Wang et. al. 2015	K562	CML	CRSIPR-KO Score
Wang et. al. 2015	Jiyoye	Burkitt's Lymphoma	CRSIPR-KO Score
Wang et. al. 2015	Raji	Burkitt's Lymphoma	CRSIPR-KO Score
Wang et. al. 2017	EOL	AML	CRSIPR-KO Score
Wang et. al. 2017	HEL	AML	CRSIPR-KO Score
Wang et. al. 2017	MOLM13	AML	CRSIPR-KO Score
Wang et. al. 2017	MonoMac1	AML	CRSIPR-KO Score
Wang et. al. 2017	MV411	Biphenotypic Leukemia	CRSIPR-KO Score
Wang et. al. 2017	NB4	AML	CRSIPR-KO Score
Wang et. al. 2017	OCI-AML2	AML	CRSIPR-KO Score
Wang et. al. 2017	OCI-AML3	AML	CRSIPR-KO Score
Wang et. al. 2017	OCI-AML5	AML	CRSIPR-KO Score

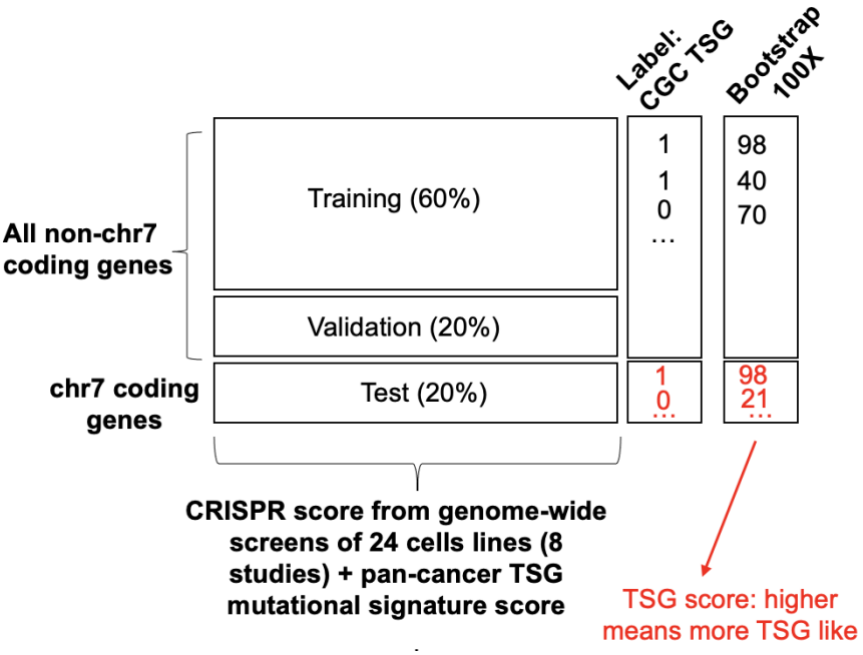
Table 2 continued

Wang et. al. 2017	P31/FUJ	AML	CRSIPR-KO Score
Wang et. al. 2017	PL21	AML	CRSIPR-KO Score
Wang et. al. 2017	SKM1	AML	CRSIPR-KO Score
Wang et. al. 2017	TF1	AML	CRSIPR-KO Score
Wang et. al. 2017	THP1	AML	CRSIPR-KO Score
Wallace et. al. 2016	MV411	Biphenotypic Leukemia (AML and ALL)	CRISPRi Score
Horlbeck et. al. 2016	K562	CML	CRISPRi Score
Horlbeck et. al. 2016	K562	CML	CRISPRa Score
Sack et. al. 2019	HMEC	Breast Cancer	ORF library overexpression proliferation Score
Sack et. al. 2019	HPNE	Pancreatic Cancer	ORF library overexpression proliferation Score
Sanson et. al. 2018	A375	Melanoma	CRISPRa Score
Sanson et. al. 2018	HT29	Colon Cancer	CRISPRa Score
Bakke et. al. 2019	PANC-1	Pancreatic Cancer	CRISPR KO Score

The "ground truth" label is a binary column specifying whether the gene is a known canonical TSG. (0 or 1, non-TSG or TSG) The data source is from Cancer Gene Census database,[213] which curates two tiers of TSGs based on the strength of literature support. Tier

1 TSGs has extensive literature support with strong experimental evidence, such as *TP53* and *PTEN*. Tier 2 TSGs has less literature support but still relatively strong evidence.[213] Since tier 2 TSGs has clear experimental evidence support, I included them and thus have 315 genes labelled as canonical TSG in total. The structure of the input data for the machine learning classifier is depicted in **Figure 3.2 A.**

A)



B)

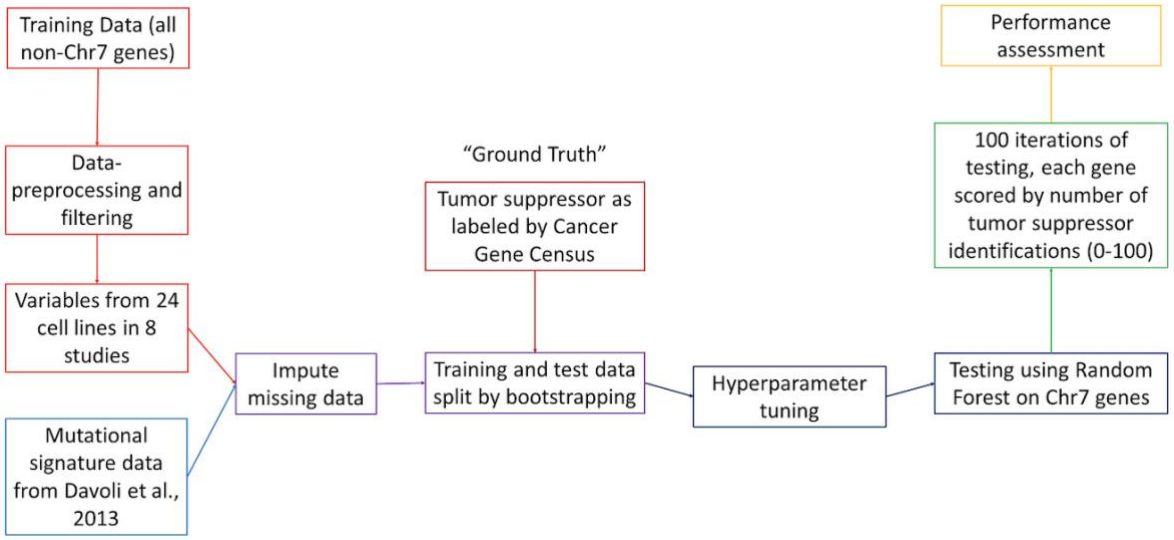


Figure 3.2 A) structure of the training, validation and testing data. **B)** Complete workflow of the machine learning model design, training, optimization and testing process.

3.2.2 Designing a robust ML workflow.

With the input data ready, next I deliberated on the proper machine learning algorithm. I focused on supervised instead of unsupervised machine learning method since it allows me to use the labelled "ground truth" TSGs to train the algorithm, in order to recognize those genes whose genome-wide screening proliferation scores behave similarly as the Cancer Gene Census labelled-TSG. The key considerations include 1) Complexity and linearity of my data and 2) The need for normalization. Since my data set include 25 distinct predictor columns and around 12,000 genes, it is expected that such complex and high dimensional dataset will display a non-linear pattern. Therefore, using a linear model such as logistic regression, or support vector machine with linear kernels will risk underfitting and over-simplifying the relationships.[272] Among the non-linear supervised machine learning models, I chose random forest because of several reasons:

1). It is easy to use with out-of-pocket implementations in R using the package caret.[273]

2). It has a good balance between accuracy and overfitting. Random forest is an ensemble method combining multiple weak learners (individual decision trees) to build a strong learner. This method usually results in strong performance than an individual learner. At the same time, random forest is good at reducing overfitting bias since each tree is training on a random set of data and features, and the final predictions are the average of multiple trees.

3). Random forest does not require data scaling or normalization since the absolute value of the training data does not affect the splitting and decisions within each decision tree (scale-invariant). And it can handle outliers well since the splitting is based on the relative

order of the data and not distance based.[274] This advantage of random forest is very important since the genome-wide screen data are obtained from different studies, performed by different scientists with different reagents and machines. An algorithm that requires normalization will introduce significant bias such as batch effects into the dataset.

Data preprocessing

Before running the learning algorithm, I performed a series of preprocessing steps to make the data ready to be input to the random forest model (**Figure 3.2B**). Even though I have removed screens with excessive missing values, our data still contain a lot of missing values, which is normal in genome-wide screen experiments as there are many drop-out (e.g. drop out gRNAs in CRISPR screen). However, missing data could be imputed by inferring the relationship with similar data points that are not missing. Among the data imputation methods, I used k-nearest-neighbor (KNN) algorithm, which estimates the value of missing data based on its nearest neighbors that are not missing. KNN method is advantageous comparing to the simple mean or median based imputation because it can preserve the distribution, capture non-linear relationships of the data and thus more accurate.[275]

After data imputation, I then split the rows (genes) into training (60%), validation (20%) and testing (20%) data. The training and validation data are all the non-chromosome 7 protein-coding gene. The training data is used for the algorithm to learn the pattern between proliferation scores and whether the gene is labelled as a TSG. The validation data is used for performance assessment, hyperparameter tuning and minimizing overfitting. The testing data are all the protein-coding genes on chromosome 7, and the test data is held out until the final prediction step. To reduce bias in training data and achieve training-testing data balance, I did 100 times of bootstrapping on training data to randomly sample from the non-chromosome 7 genes. For each bootstrap, I performed hyperparameter tuning of the random forest model

and select the combination of hyperparameters that gives the smallest out-of-bag (OOB) error, using a grid search approach. The hyperparameter I tuned include number of decision trees, tree max depth, minimal sample required to split a node leaf and maximum number of features considers for a node. I chose gini impurity as the function to define the quality of each node split. I then performed prediction on the testing data using the 100 tuned models and obtained a binary result for each gene (1 for putative tumor suppressor and 0 for non-tumor suppressor). I then devised a "TSG score" by ranking the chromosome 7 genes based on the frequency of being labelled as tumor suppressors in the 100 iterations. The higher the score is, the more it behaves like a TSG. For example, if a gene is labelled as TSG in 60 out of 100 bootstrapped model testing, its TSG score is 60/100 = 0.6.

I achieved an area under the curve (AUC) of 0.777 (95% CI 0.747-0.806), as seen in the receiver operating characteristic (ROC) curve in **Figure 3.3A.** I achieved a mean out-of-bag error rate of 0.29 (**Figure 3.3B**), and an average of 71.8% accuracy, 73.4% precision, 59.5% sensitivity, and 82.0% specificity. In order to decide which feature (screen data) gives the best performance, I ranked the mean decrease in gini index by removing each feature from the model individually (**Figure 3.3C**). A higher decrease in gini index means removing the feature leads to a less pure decision node and thus worse classification result, indicating this feature is more important in giving the correct predictions. TSG-specific mutational signature came on top as the most powerful features in predicting TSGs. This supports the results from Davoli et. al 2013.[42] The genome-wide perturbations screens have relatively comparable decrease in gini index. The feature importance is orthogonally validated by mean decrease in accuracy, which returns similar ranking as gini index (**Figure 3.3D**). In summary, the random forest model returns high accuracy for predicting TSGs based on the ground truth labelled by cancer gene census. Next, it will be valuable to cross check the ranked ML predicted TSGs with experimental results.

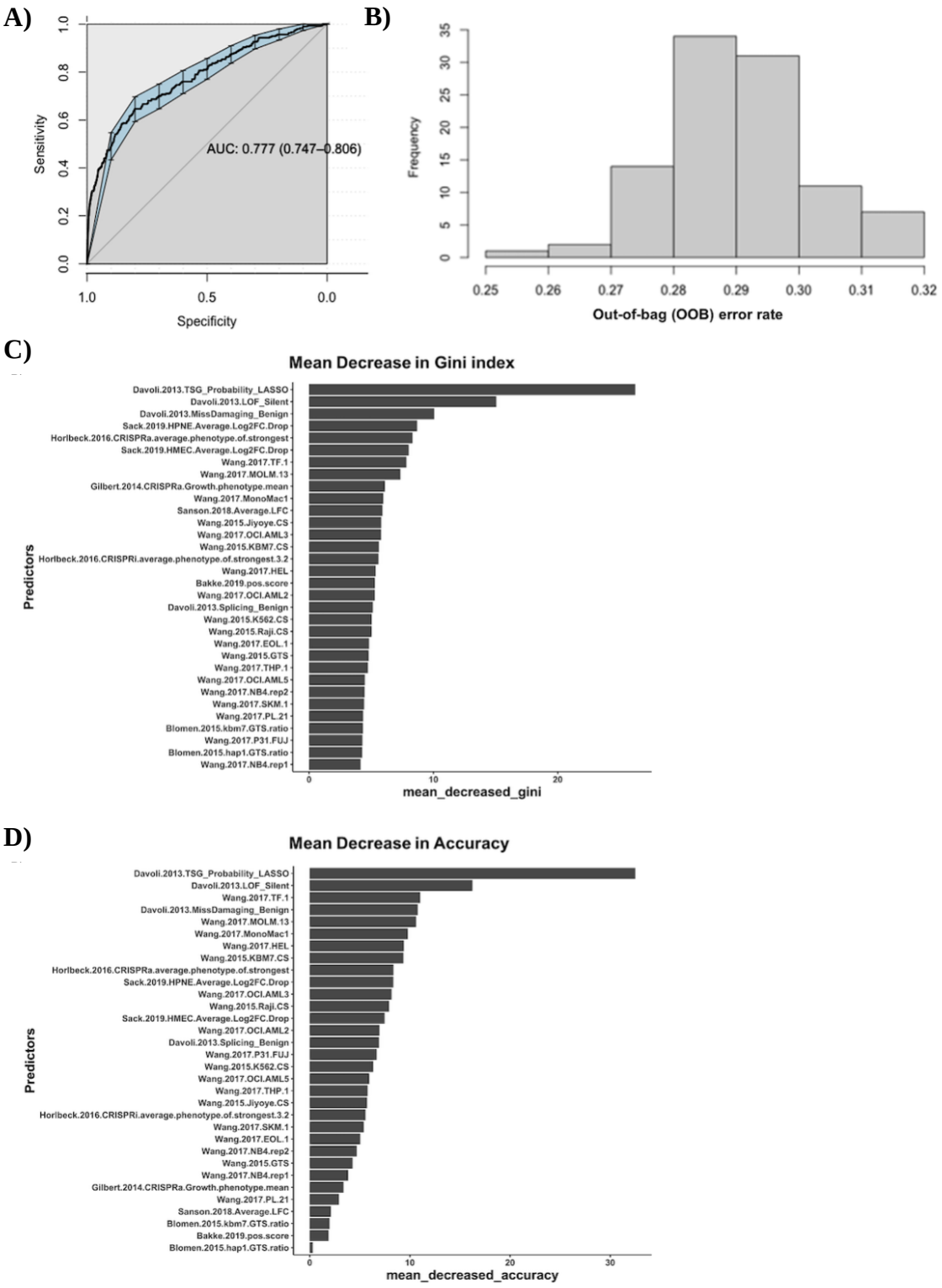


Figure 3.3 Quality control and performance measures for machine learning model. (**A**) AUC (area under the curve) ROC (receiver operating characteristics) curve. Error bar represents 90% CI. (**B**) Out-of-bag (OOB) error rate frequency distribution, representing the distribution of OOB error rate of 100 bootstraps of the random forest model. mean = 0.290, n=100. Represents the average error for each iteration using predictions from the trees that do not contain it within their respective bootstrap sample. (**C,D**) Importance of classification

Figure 3.3 continued variables across 100 bootstrap iterations. Each classification variable is defined by the technology, cell line/mutation signature, and reference (Author,year)

3.2.3 ML cross-validated the experiment in identifying chr7 TSGs

In parallel to the *in silico* machine learning approach, my collaborator Jeremy Baeten designed and implemented an *in vitro* plate array based CRISPR KO screen.[214] For the screen, he compiled a list of 161 genes based on the following criteria: 1) They are hits from previous genome-wide screens. (For example, knocking out of the gene increased cancer cell line proliferation). 2). They are labeled as TSG from Cancer Gene Census database.[213] 3). They possess TSG-specific mutational patterns as identified by Davoli et. al. 2013.[42] 4). They are expressed in human CD34+ HSPC. The screen is performed in the human primary HSPC, since genetic dysregulation in this population is the direct cause of human myeloid malignancies. 5). All chromosome 7 commonly deleted region genes.

gRNAs were designed for each gene, and 108/161 gRNAs passed editing efficiency threshold.[214] CD34+ HSPCs are transfected with Cas9-gRNA ribonucleoproteins, so that each well contains one gRNA targeting one gene. Two parallel screens including proliferation in maintenance media and erythroid differentiation assays in EPO-containing media are performed to assay the TSG properties. The underlying assumption is that knocking out a TSG will lead to impaired erythroid differentiation and increased HSPC proliferation. Proliferation and erythroid differentiation scores are calculated as the normalized mean proliferation measurement for each gRNA transfected well comparing to gAAVS1 (control). gPTEN is used as the positive control for increased proliferation, and gGATA1 is included as the positive control for impaired erythroid differentiation.

Combining the proliferation & erythroid differentiation assay results will give a more complete measurement of TSG properties, rather than looking at them individually. However, the readouts of the two assays are in different units. Therefore, I devised a schematic to unite the proliferation and erythroid differentiation scores into a combined score that reflect the tumor suppressor property of the target gene: for both erythrocyte differentiation and proliferation scores, I normalized the experimental results of each gene in both assays in each well by subtracting from AAVS1 control. Erythroid signs were inverted so that a higher score is associated with increased proliferation and decreased erythroid differentiation. Then I obtained an average erythrocyte differentiation and proliferation score by taking the mean of each gene's results across all replicate plates. I then combined the average erythrocyte differentiation and proliferation scores into one table and performed min-max normalization onto a 0-1 range, in order to remove the effect of directionality and unify the two scores onto the same scale. Finally, I summed the normalized proliferation and erythrocyte differentiation score to obtain the "combined experimental score", which is a unified measure of the likeliness for each gene to be a tumor suppressor. To test whether the proliferation and erythrocyte differentiation results for each gene are significantly different from those of the AAVS1 control, I performed a non-parametric Mann-Whitney test on the results of each gene across all replicate wells compared to AAVS1. The p values were multiple hypothesis corrected using Storey's q value (FDR < 0.12). The combined score is shown in **Figure 3.4A**. Across all candidate genes, increased proliferation correlates with impaired erythropoiesis, consistent with the known link between these two phenotypes. The positive control for proliferation gPTEN and for erythroid differentiation gGATA1 are among the top significant genes with highest combined TSG score, supporting the robustness of the CRISPR screen. 12 genes whose knockout lead to significant greater proliferation and more impaired erythroid differentiation comparing to the gAAVS1 control are highlighted in red. Some of these genes

have been shown to possess TSG properties, such as TRIM24,[276] while others are less well characterized. To compare the results from CRISPR screen and the machine learning classifier, z scores of the machine learning score, experimental proliferation, erythroid differentiation impairment and the combined scores are shown side-by-side in **Figure 3.4B.** Many highly scored genes overlapped with genes that scored significantly experimentally, such as *CUX1*, *LUC7L2* and *TRIM24*.[276,277] The overlap of the experimental and classifier results did not reach significance (hypergeometric test $p = 0.12$). Conceivably, this may reflect a limitation of the classifier and/or some hits in the classifier may exhibit tumor suppressor activity by other measurements, such as apoptosis, metastasis, or DNA repair. Nonetheless, using the classifier scores, genes within CDRs are again significantly enriched for TSGs (**Fig. 3.4C, D**). This result from disparate datasets, across tumor types, mirrors our experimental results. At the time of the manuscript writing, the successful application of machine learning with genomic and CRISPR screen data to identify TSGs has not been previously reported. Furthermore, our result buttresses the concept of CDRs manifesting as a contiguous gene syndrome.

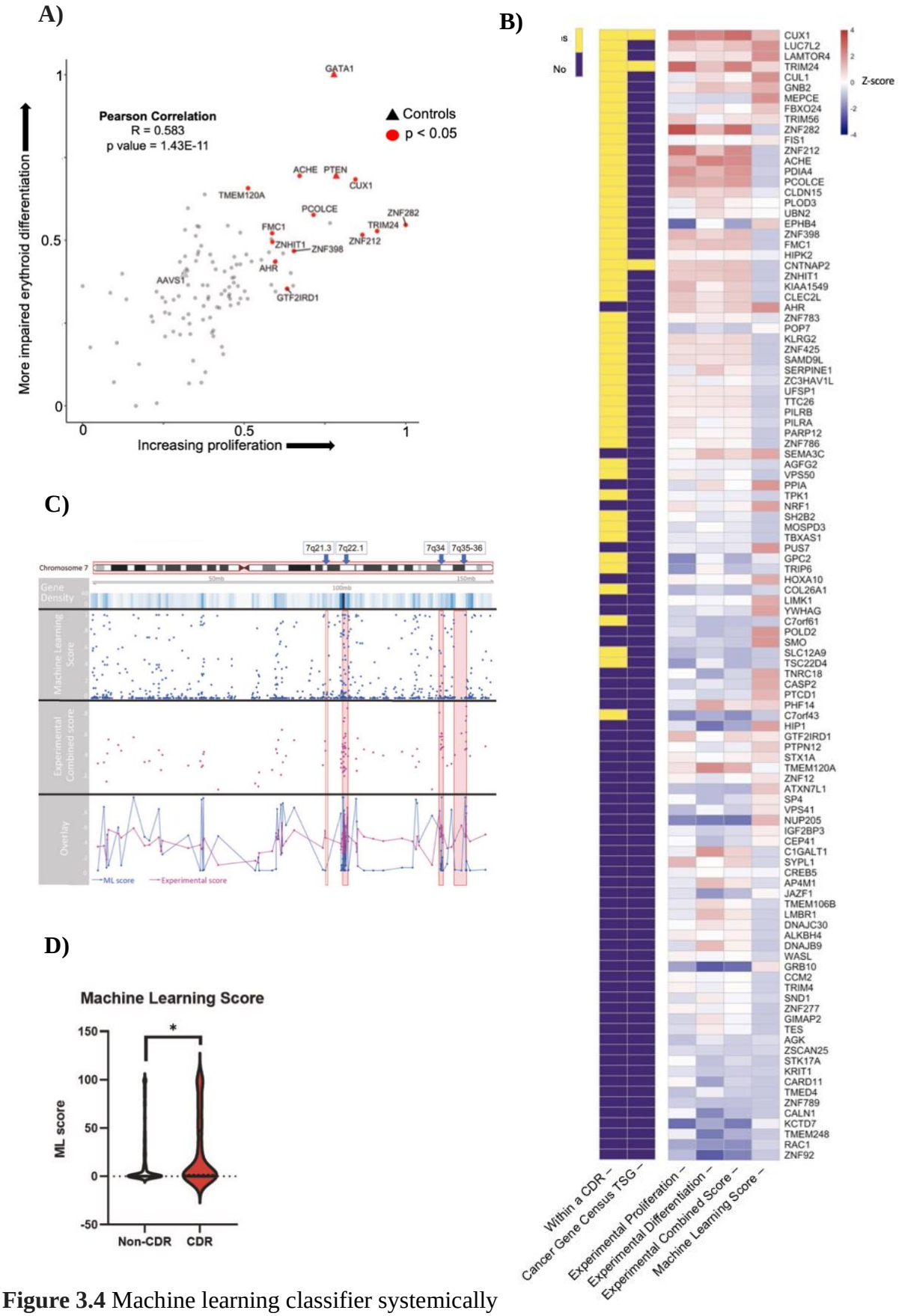


Figure 3.4 Machine learning classifier systemically ranked chromosome 7 gene TSG-like activities and cross-validated *in vitro* CRISPR screen

Figure 3.4 continued experiment **A)** Combined score of the 108 candidate TSGs validated by CRISPR screen. x and y axis are transformed proliferation and erythroid differentiation score, respectively. **B)** Heat map of variables used to consider myeloid TSG status in all genes included in experimental analysis. Genes are ranked by experimental combined score, machine learning score, and CDR status equally weighted. Columns 1 and 2 are binary variables where yellow = yes and purple = no; the remaining columns are z-scores of the experimental and machine learning classification variables. **C)** Genomic track of all chromosome 7 genes. Rows depict gene density, machine learning score, combined proliferation and erythroid differentiation score, and overlay of machine learning (ML) and experimental scores. Red boxes indicate CDRs. **D)** Machine learning score of genes within (n = 74) or outside of CDRs (n = 825). Significance determined by Mann–Whitney–Wilcoxon test, *$p < 0.05$.

3.3 DISCUSSION

The application of machine learning in identifying the mechanistic link between genetic functions and cancer development is a field developing at warp speed, driven by the ever more abundant multimodal data including RNA-seq, single cell RNA-seq, epigenomics data and clinical data. This project leveraged machine learning to unravel the link between genetic functions and cancer development, focusing on tumor suppressor genes (TSGs). Comparing to the conventional manual process of identifying the statistically significant hits in each perturbation screen and compile them, a key advantage of this project is that ML could identify the subtle but consistent changes across all input datasets and capture the hits that might be otherwise omitted by the thresholded manual analysis approach. By systematically mining publicly available genome-wide perturbation screens, the project provided a framework to discover tumor suppressor properties at a large scale by learning the behaviors of known TSGs from the phenotype of cellular proliferation.

However, tumor suppressor function is not just driven by cellular proliferation, rather, dysregulation of TSGs in cancer is a multi-dimensional process involving apoptosis, metastasis, metabolism, and DNA repair.[278] These processes might or might not contribute to cell proliferation manifested in perturbation screens. Some of the top hits such as *CUX1* and

TRIM24 are consistently ranked as a TSG by both the experimental CRISPR screen and the ML classifier. However, there are also some genes possessing TSG activity in the in vitro assay, but are ranked as low possibility TSGs by the ML classifier. (**Figure 3.4B**). This discrepancy might be partially explained by the fact that the experimentally determined TSG could exert their activities on aspects beyond cellular proliferation. Ideally, a better classifier would cover the multifaceted mechanism of actions of TSGs by incorporating multimodal data sources such as imaging, epigenomics structural proteomics and metabolomics datasets. Studies in recently years have already utilized these dataset to predict tumor suppressors.[279–282] The effective integration of these multi-modal data and precise stratification based on patient population heterogeneity will present exciting opportunities to translate these AI/ML approaches to clinical applications.

My approach aligns well with the recent shift towards understanding the complex nature of cancer and the intricate interplay of various cellular processes in tumorigenesis. Machine learning models, when trained with comprehensive and diverse datasets, can uncover subtle patterns and interactions that are often elusive to conventional analysis. However, the principle of "garbage in, garbage out" highlights the importance of the quality and relevance of the input data. There are multiple ways this classifier could be improved in the future. First, it could be iteratively improved by incorporating more hematopoietic genome-wide screen data since the date of publishment. Furthermore, TSGs could function as oncogenes in different context such as cancer type, cell type and stages of cancers.[283,284] Even within hematological malignancies, the same gene could function to either promote or inhibit tumorigenesis. For example, *RUNX1* is frequently mutated in myeloid neoplasms including MDS and cytogenetically normal AML, and is widely considered as a tumor suppressor.[285] However, wild type *RUNX1* has also been shown to show to promote myeloid

leukemogenesis in MLL fusion AML.[286] Therefore, the training data should be stratified based on existing knowledge on the cancer context.

This project, at the time of ideation, is one of the early efforts to discover tumor suppressor properties in a large scale by systemically mining publicly available genome-wide perturbation screens. The classifier provided a concise framework to learn the behaviors of TSGs from the phenotype of cellular proliferation.

CHAPTER 4: CUX1 REGULATES HUMAN HEMATOPOIETIC STEM CELL CHROMATIN ACCESSIBILITY VIA THE BAF COMPLEX

4.1 INTRODUCTION

Multipotent stem cells are crucial for adult tissue maintenance and function. Disruptions in their homeostasis and lineage commitment lead to various diseases, including cancer.[85] Understanding lineage determination is a key developmental biology question, critical for creating therapies that target stem cell dysfunction.

Cell fate is shaped by various external signals and internal factors such as epigenetic regulators and TFs. TFs coordinate signalling cues to direct genomic reprogramming for cell type specific gene expression.[91] TFs modulate transcription by recruiting proteins that physically remodel nucleosomes, modify histone and DNA, or regulating RNA polymerase directly.[86] Pioneer TFs promote DNA accessibility for subsequent TF binding by recruiting nucleosome remodeling enzymes such as the SWI/SNF (or BAF, BRG/BRM-associated factor) complex, are crucial in the cell fate determination process by promoting lineage-restricted gene expression programs.[112]

The ATPase dependent BAF complexes, consisting of 10-13 subunits, plays important role for various biological processes including transcription, DNA repair, and development.[137] They reconfigure nucleosomes for gene expression pertinent to lineage differentiation.[136] The BAF complex lacks DNA binding domains and relies on TFs for DNA targeting, as seen in hematopoiesis with TFs like RUNX1, PU.1, and KLF1.[157,158,287] Nonetheless, these TFs only account for a portion of BAF chromatin binding, implicating additional, yet unknown, hematopoietic pioneer TFs.

CUX1 is a homeodomain-containing TF essential for various cellular functions, including neural, lung, and hematopoietic tissue differentiation.[75,79,166,167] Mutations in CUX1

are linked to developmental delays and cancers.[69,288] It plays essential roles in regulating HSC homeostasis, lineage determination, and acts as a tumor suppressor.[36,79] CUX1 binds DNA through three CUT repeats and one homeodomain. CUX1 binding is enriched at enhancers, particularly those in active contact with promoters, suggesting it influences gene expression over long distance looping.[169,289] Its role in transcription is context-dependent, with capabilities to activate or repress gene expression by mechanisms including competing for DNA binding sites, and interacting with other TFs, chromatin modifiers and transcription co-activator.[67,290–293] Together, these data indicate that CUX1 is an epigenetic modifier that interfaces with higher order chromatin structure, yet the molecular mechanism by which CUX1 controls transcription is incompletely understood. In this study, we address this question by identifying endogenous CUX1 interacting partners, CUX1 genomic targets, and the ensuing epigenetic consequences through unbiased proteomics and genome-wide functional genomics approaches in a human leukemia cell line and primary human HSPCs.

4.2 RESULTS

4.2.1 CUX1 recruits the BAF chromatin remodeling complex to enhancers.

To determine the mechanism by which CUX1 governs gene expression, we identified CUX1 protein interaction partners by performing co-immunoprecipitation for endogenous CUX1 followed by mass spectrometry in the K562 human myeloid leukemia cell line. We chose K562 cells for this experiment as they are considered human leukemic representatives of multipotent progenitors, capable of differentiation into erythroid, megakaryocytic, and myeloid lineages.[294–296] This analysis revealed nine components of the BAF complex interacting with CUX1 (FDR<0.05) (**Figure 4.1A**). Many of the protein subunits identified

are shared across the three major BAF complexes; however, the detection of ARID1A and ARID1B suggests that CUX1 interacts with the canonical BAF complex (cBAF).[153] CUX1 interactions with two core BAF complex members, SMARCA4 (BRG1) and SMARCC1 (BAF155) were confirmed by western blot (**Figure 4.1B**).

We next tested if CUX1 and BAF bind to overlapping genomic loci. We performed ChIP-seq for CUX1 and SMARCA4, the essential enzymatic BAF subunit.[297] Using the thresholded peak-calling method by MACS2 and IDR analysis,[221,222] in total 66.4% (17,595/26,497) of CUX1 binding sites overlapped with SMARCA4 peaks, revealing extensive overlap of CUX1 and SMARCA4 on DNA (**Figure 4.1C**). CUX1 and SMARCA4 overlapping sites were localized predominantly at enhancers (**Figure 4.1E**) and enriched for the hematopoietic TF motifs GATA, RUNX1, KLF1, and PU.1 (**Figure 4.1D**). Significant overlap of CUX1/SMARCA4 co-bound sites with published ChIP-seq data,[298,299] shows that CUX1 and the BAF complex interact with other hematopoietic TFs at enhancers (**Figure 4.1F**).

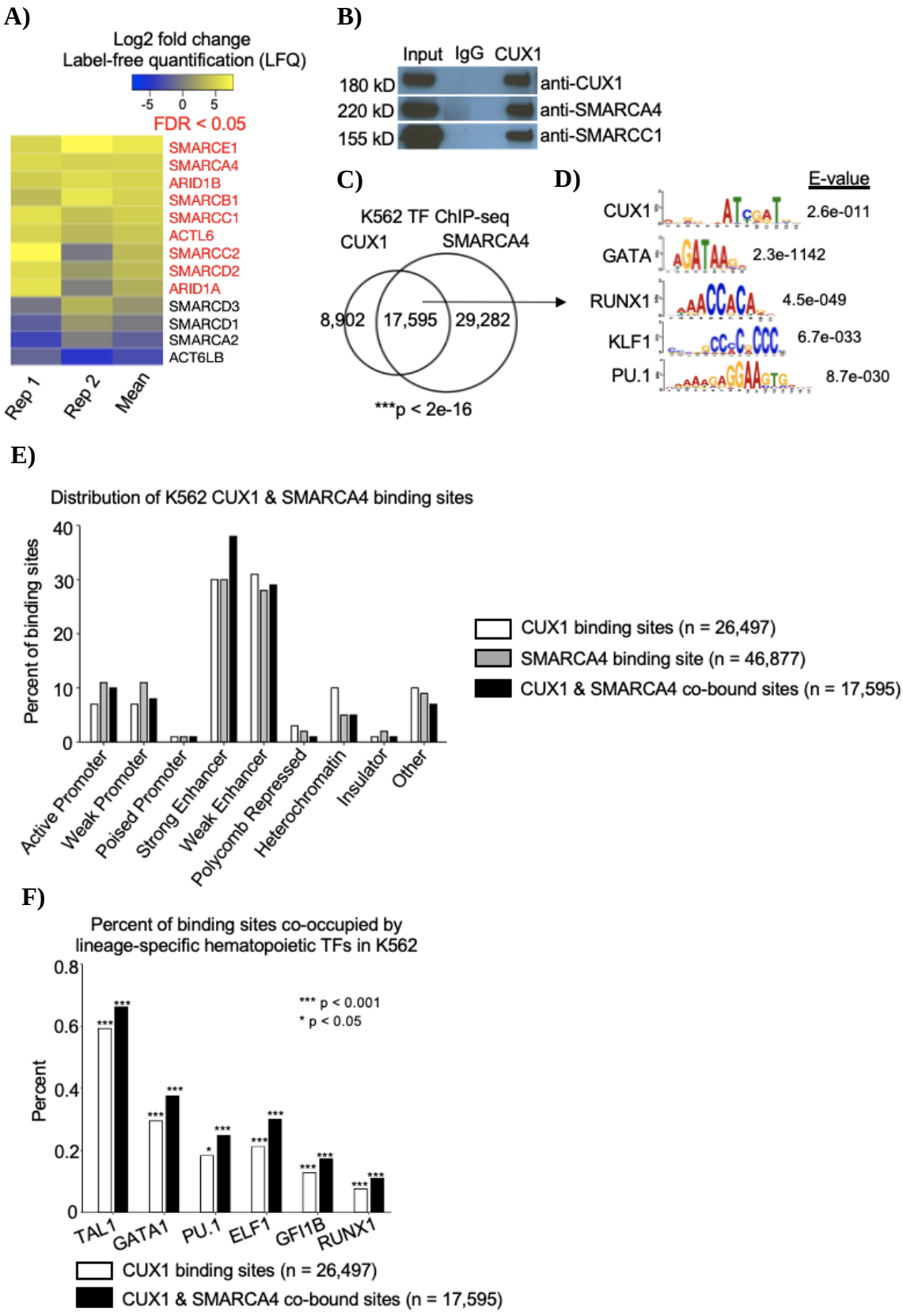


Figure 4.1: CUX1 co-occupies genomic loci with the chromatin remodeler BAF complex **A)** Co-immunoprecipitation for CUX1 in K562 cells was followed by mass spectrometry (n=2 biological replicates). The heatmap indicates BAF members ranked by the mean label-free

Figure 4.1 continued quantification fold enrichment compared to IgG controls. Red indicates FDR<0.05. **B)** Representative co-immunoprecipitation followed by immunoblot in K562 (n=2 biological replicates). **C)** K562 CUX1 and SMARCA4 ChIP-seq overlap (n=2 biological replicates, IDR<0.05). **D)** Enriched motifs92 at CUX1 and SMARCA4 co-occupied sites. **E)** Chromatin state annotation of all CUX1 binding sites (n=26,497), all SMARCA4 binding sites (n=46,877) and CUX1 and SMARCA4 overlapping sites (n=17,595) via intersecting binding sites with the K562 ChromHMM track. **F)** Percent of CUX1 binding sites and CUX1/SMARCA4 co-bound sites from ChIP-seq co-occupied by lineage specific hematopoietic TFs (all pairwise comparisons of overlap are significant by hypergeometric test, p<0.05). All hematopoietic TF ChIP-seq binding sites are obtained from ENCODE database.

We next tested the hypothesis that CUX1 recruits BAF to DNA. We performed ChIP-seq for SMARCA4 in K562 clones CRISPR/Cas9 edited for CUX1 (gCUX1) or a control intronic region of HPRT (gHPRT).[81] Among the 49,070 (IDR<0.05)[222] SMARCA4 binding sites identified in gHPRT control cells, 52.1% (25,565) were reduced in gCUX1 cells (CUX1-dependent SMARCA4 sites) (**Figure 4.2A**). An example of the reduction of SMARCA4 binding after CUX1 knockout is shown at the HMBS gene, encoding the essential erythrocyte hydroxymethylbilane synthase enzyme (**Figure 4.2E**).[300] This experiment shows that CUX1 promotes recruitment of the BAF complex to bind certain loci.

Next, we interrogated if CUX1 directly recruits SMARCA4. An intersection of CUX1-dependent SMARCA4 sites with CUX1 binding sites revealed that 27.4% (7,005/25,565) of CUX1-dependent SMARCA4 sites are at loci directly bound by CUX1 (**Figure 4.2B, left**). In this 'direct model', CUX1 promotes recruitment of SMARCA4 to a substantial fraction of DNA binding sites. We next examined CUX1 binding at the CUX1-independent SMARCA4 sites. To this end, we intersected CUX1-independent SMARCA4 sites with CUX1 ChIP-seq peaks. 41% (9,850/23,505) of these sites were bound to CUX1 (**Figure 4.2B, right**). This finding suggests that while CUX1 is not necessary for SMARCA4 binding at these loci, SMARCA4 may still be co-bound with CUX1, referred herein as an 'indirect model' of SMARCA4 binding. These 7,005 and 9,850 sites are referred to hereafter

as “CUX1 recruited SMARCA4” and “non-CUX1 recruited SMARCA4” sites, respectively. Example genome snapshots of these two categories are shown in **Figure 4.2C** (Left: KLF1, encoding a TF essential for erythropoiesis.[301,302] Right: ENDOD1, encoding a nucleic acid hydrolyzation nuclease).[303] To understand the differences between these two categories, we further characterized the underlying features of these sites. While 69.1% of non-CUX1 recruited SMARCA4 sites are at distal regulatory elements, this increases to 84% for CUX1 recruited SMARCA4 sites, suggesting that CUX1 recruits SMARCA4 to many distal enhancers (**Figure 4.2D**). Further, differential motif analysis shows that the CUX1-recruited SMARCA4 sites are enriched for lineage-specifying TFs (**Figure 4.2F**). These data are compatible with a model where CUX1 promotes BAF recruitment, particularly at enhancers potentially regulated by lineage-directing TFs.

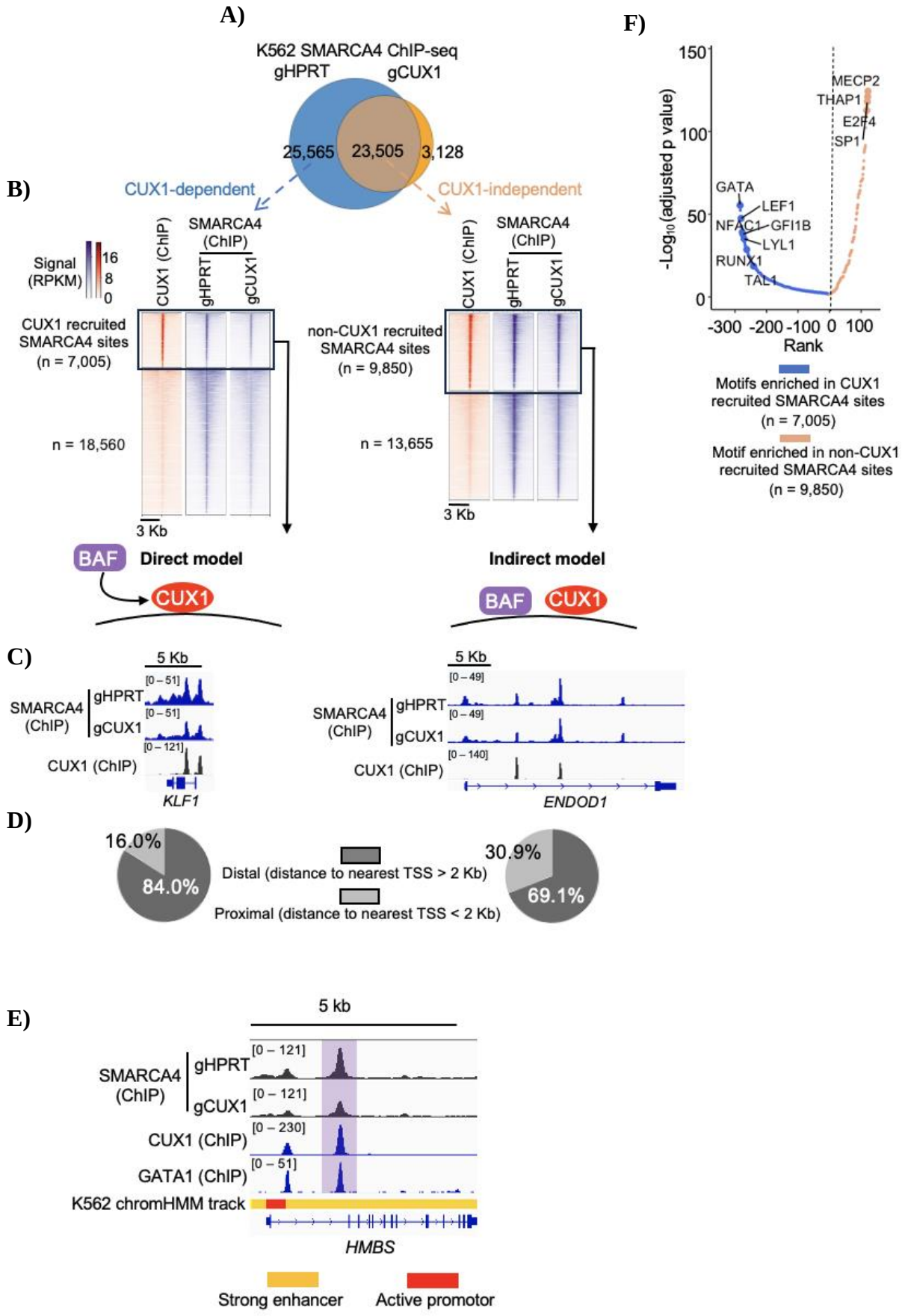


Figure 4.2: CUX1 recruits the BAF chromatin remodeling complex to enhancers.

Figure 4.2 continued A) Overlap of SMARCA4 peaks (n=2 biological replicates, IDR<0.05) in gHPRT and gCUX1 K562 cells. **B)** Heatmaps showing overlap between CUX1-dependent or CUX1-independent SMARCA4 sites with CUX1. The values are normalized ChIP-seq reads (RPKM). The direct model represents CUX1 recruitment of SMARCA4. The indirect model represents SMARCA4 sites bound but not recruited by CUX1. Example genome snapshots for each category are shown **C)**.[225] **D)** Distance to the nearest transcription start site (TSS) of CUX1-recruited and non-CUX1-recruited SMARCA4 sites. **E)** IGV analysis of ChIP-seq tracks at the HMBS erythroid gene. Tracks shown are normalized ChIP-seq signal across 2 replicates (RPKM) for K562 SMARCA4 gHPRT, gCUX1, CUX1 and GATA1, along with K562 chromHMM chromatin state annotations. **F)** Differential motif analysis of CUX1-recruited vs.non-CUX1-recruited SMARCA4 sites

4.2.2 CUX1 with SMARCA4 promotes the establishment of accessible chromatin.

As recruitment of BAF is one mechanism through which pioneer transcription factors remodel chromatin,[112] we next assessed the role of CUX1 in the regulation of DNA accessibility. We performed ATAC-seq[218] on gCUX1 and control gHPRT K562 cells. To investigate the effect of CUX1 on DNA accessibility, we first applied a non-thresholded, quantitative approach. To this end, we performed genome-wide differential accessibility analysis using csaw[232] on the ATAC-seq data and observed more sites with significantly downregulated (n=933) than upregulated (n=210) accessibility after loss of CUX1 (FDR<0.05, |Log2FC|>1) (**Figure 4.3A**), indicating that CUX1 normally contributes to chromatin opening. Among the 933 significantly decreased ATAC sites, a considerable proportion (38.1%) are at enhancers (**Figure 4.3B**) and are enriched for PU.1 and KLF1 motifs (**Figure 4.3A**), consistent with the model that CUX1 promotes chromatin accessibility at enhancers involved in hematopoietic differentiation. The changes in accessibility were accompanied by concordant changes in the activating chromatin mark H3K27ac, indicating that CUX1 maintains enhancer activation and accessibility in K562 cells (**Figure 4.3C**).

Normalized ChIP-seq reads of both CUX1 and SMARCA4 are significantly higher at the significant down ATAC sites (933 peaks) compared to the CUX1-independent ATAC sites (14,256 peaks with the least significant change in chromatin accessibility. FDR>0.05,

|Log2FC|<0.5) (**Figure 4.3D**). This finding that sites normally opened by CUX1 have higher occupancy of both CUX1 and SMARCA4 suggests a direct involvement of these factors in driving chromatin accessibility. Analysis of all sites co-bound by CUX1 and SMARCA4 demonstrated a significant drop in accessibility after CUX1 knockout (p<2.2e-16) (**Figure 4.3E**); CUX1-knockout decreased accessibility at CUX1-recruited SMARCA4 sites (**Figure 4.3F**) consistent with a model in which CUX1 recruits BAF to enhancers and increases DNA accessibility. Unexpectedly, CUX1 also influences accessibility independent of its ability to directly recruit the BAF complex. (i.e., the non-CUX1 recruited SMARCA4 sites) (**Figure 4.3G**). While not tested here, this later finding may be due to CUX1 recruitment of additional activating factors, such as HATs,[292,293] or downstream indirect effects of CUX1 loss.

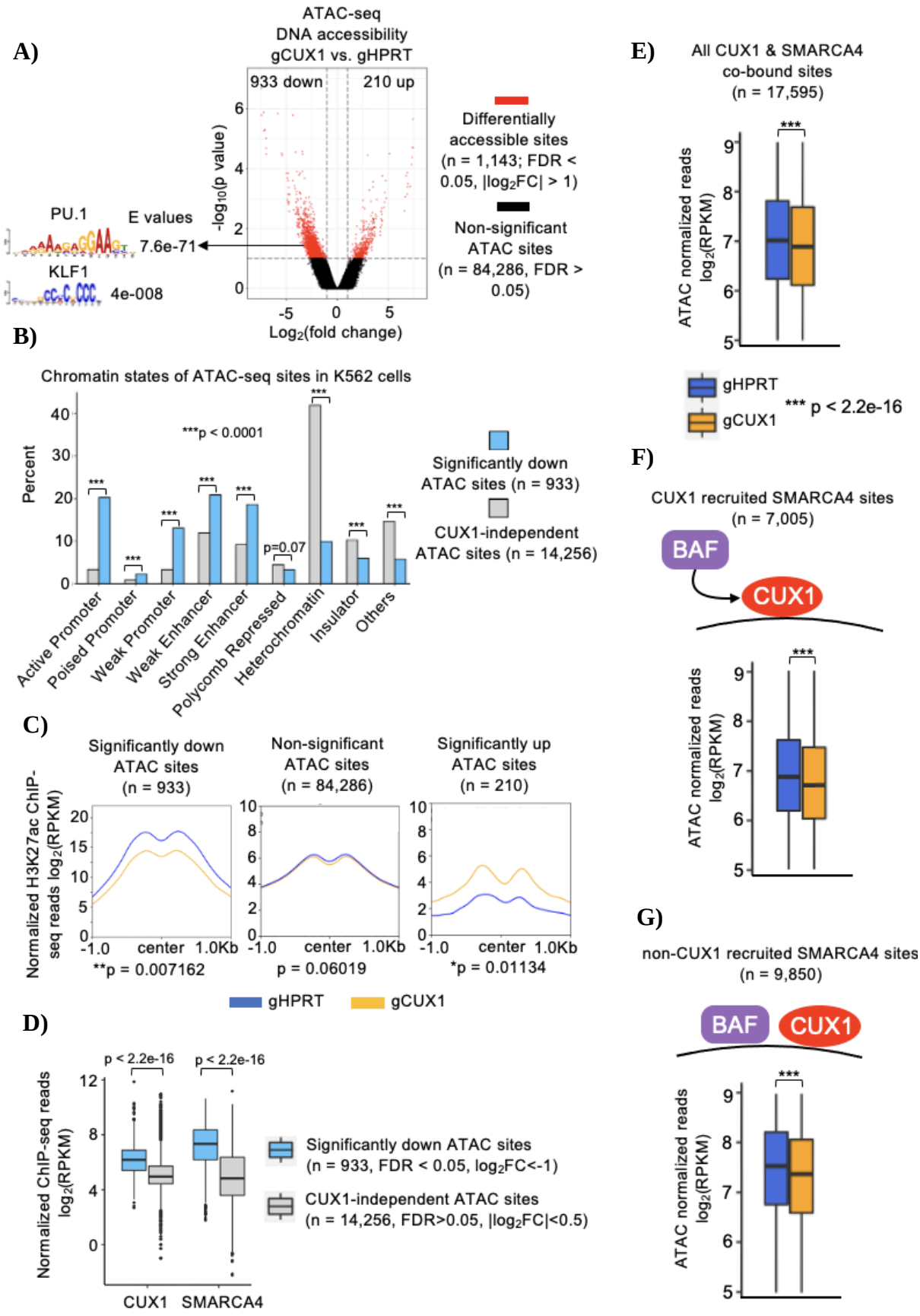


Figure 4.3: CUX1 with SMARCA4 promotes the establishment of accessible chromatin.

Figure 4.3 continued A) Volcano plot comparing ATAC-seq signal in gCUX1 vs. gHPRT K562 cells (n=2 biological replicates). Significance calculated by csaw.[232] Top enriched motifs for the significant down sites are shown. **B)** Distribution of K562 chromHMM chromatin state of the sites whose accessibility are significantly downregulated after CUX1 loss (n=933, blue, FDR<0.05, log2FC<-1) and the ATAC sites whose accessibility are not dependent on CUX1(n=14,256, grey, FDR>0.1, |log2FC|<0.5), significance calculated by hypergeometric test. **C)** H3K27ac ChIP-seq reads (n=2 biological replicates) at significantly down, up and non-significant ATAC sites. **D)** CUX1 and SMARCA4 occupancy at down (n=933) vs. CUX1-independent ATAC sites (n=14,256). ATAC-seq signal from gHPRT and gCUX1 cells for CUX1 and SMARCA4 co-occupied sites **E)**, CUX1-recruited SMARCA4 sites **F)**, and SMARCA4 sites bound but not recruited by CUX1 **G)**. Significance for (C-G) calculated by two-sided Wilcoxon rank-sum test.

4.2.3 In human HSPCs, CUX1 maintains DNA accessibility at enhancers associated with SMARCA4 and hematopoietic differentiation.

To observe whether CUX1 co-occupies genomic loci with BAF components in primary human CD34+ HSPCs, we used CUT&RUN in lieu of ChIP-seq as CUT&RUN requires fewer cells.[217] We observed that 52.1% (3,519/6,758) of CUX1 binding sites overlap those of SMARCA4 (**Figure 4.4A**), and CUX1 and SMARCA4 binding signals are highly correlated with each other genome-wide (Spearman's ρ=0.71, p<2.2e-16) (**Figure 4.4B**). Compared to ChIP-seq in K562 cells, CUT&RUN in CD34+ cells showed a relative enrichment for CUX1 and SMARCA4 at promoter-proximal binding sites (**Figure 4.4C**). It is unclear if this shift is due to technical differences in the assays or biological differences between the cell types. Nonetheless, CUX1 and SMARCA4 binding signals remain positively correlated with activating chromatin marks in HSPCs from the NIH Roadmap Epigenomics database, with correspondingly higher correlations with H3K4me3, associated with promoters (**Figure 4.4B**). The CUX1/SMARCA4 co-bound sites at promoter proximal (n=2,444) and distal (n=1,075) regions were assigned to the single nearest gene using GREAT and functionally annotated using AMIGO.[226,236,304–306] Notably, the distal genes were enriched for processes involved in cellular differentiation and morphogenesis. In comparison, the proximal genes were enriched for more general cellular processes such as transcription

and mitosis (**Figure 4.4D**). Examples of CUX1 and SMARCA4 co-occupancy at enhancers of genes important for multilineage hematopoietic cell differentiation, GFI1B and RUNX1,[102,307–309] and at promoters of the mitosis and DNA transcription related genes, TUBB and MED18, are shown (**Figure 4.4E, 4.4F**).[310,311]

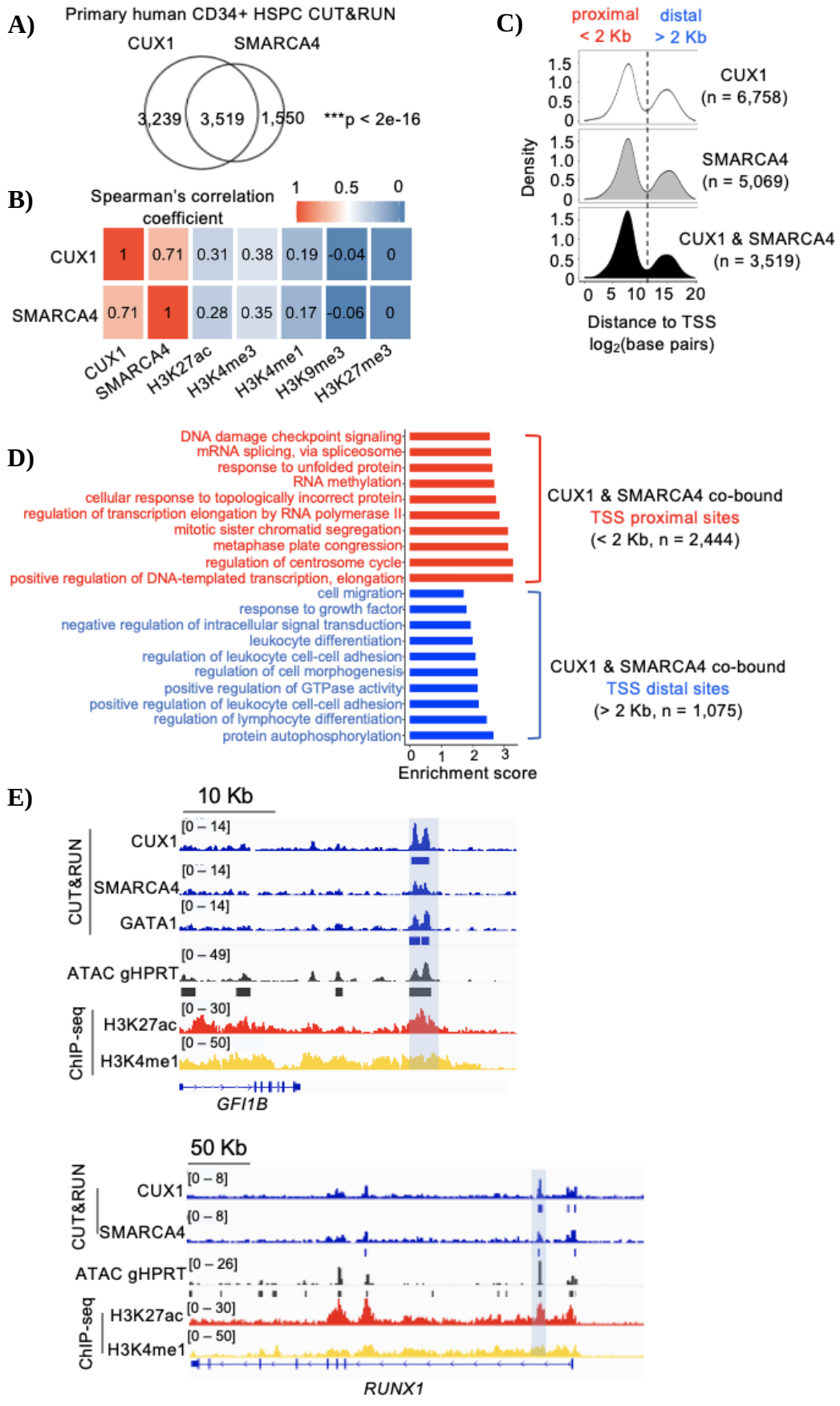


Figure 4.4: CUX1 and the BAF complex co-occupy genomic loci in primary human HSPC.

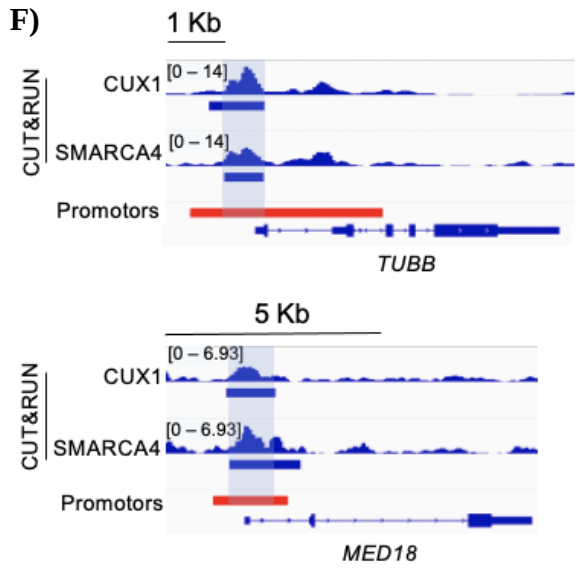


Figure 4.4 continued A) Overlap of CUX1 and SMARCA4 CUT&RUN peaks in primary human CD34+ HSPCs (n=2 biological replicates, IDR<0.05). **B)** Genome-wide correlation of CUX1 and SMARCA4 CUT&RUN signals with histone marks from Roadmap Epigenomics.[236] All pairwise correlations have p<0.001. **C)** CUX1 and SMARCA4 peaks absolute distance (log2 transformed) to the nearest TSS. The dash line indicates 2 Kb. **D)** Top GO terms for TSS-proximal and -distal CUX1/SMARCA4 co-bound sites (Bonferroni corrected p-value<0.05).[226,304] IGV genome snapshots of human CD34+ HSPC CUT&RUN data showing CUX1 and SMARCA4 co-occupancy at **E)** enhancers of hematopoietic lineage-specifying genes GFI1B and RUNX1. ATAC-seq gHPRT control track and active enhancer-specific histone modifications H3K27ac and H3K4me1 tracks obtained from Roadmap Epigenomics are added. **F)** CUX1 and SMARCA4 co-occupancy at promoters of essential genes involved in mitosis TUBB, and MED18, which is a subunit of the mediator complex that is essential in DNA transcription. Promoter annotation labelled by Roadmap Epigenomics are added. Highlighted areas are CUX1 and SMARCA4 peaks called by MACS2 (solid rectangles, IDR < 0.05)

To assay accessibility following CUX1 loss in primary cells, we transfected human CD34+ HSPCs with CRISPR gRNAs targeting HPRT and CUX1 for ATAC-seq analysis 48 hours post-transfection. The mean editing efficiency of CUX1 was 49.5% and 75.5% for HPRT. Differential accessibility analysis using csaw[232] showed that 1,603 sites were significantly lost (FDR<0.05, log2(FC)<-1) and only 3 were gained (FDR<0.05, log2(FC)>1) after CUX1 editing, confirming that CUX1 promotes open DNA accessibility in primary HSPCs (**Figure 4.5A**). Most of the significantly lost ATAC-seq sites are located at predicted enhancers (**Figure 4.5B**) and show an enrichment of multiple hematopoietic TF motifs including PU.1, RUNX1, C/EBPα, TAL1, and HLF (**Figure 4.5A**). These TFs play key roles in lineage commitment and maintaining hematopoietic stem cell (HSC) quiescence.[122,312] To further quantify the effect of CUX1 on enhancer accessibility, we obtained 3,902 genome-wide CUX1-bound enhancers by intersecting CUX1 binding sites from CUT&RUN and the human CD34+ chromHMM track from the NIH Roadmap Epigenomics database.[236] Enhancers bound by CUX1 have significantly greater DNA accessibility than enhancers not bound by CUX1 (**Figure 4.5C**). Next, we focused on the enhancers directly bound by CUX1 (n=3,902) and observed that upon CUX1 loss, there is a significant decrease of accessibility (**Figure 4.5C**), indicating CUX1 is required to promote open chromatin at enhancer regions. Lastly, to examine the relationship of CUX1-mediated accessibility with the BAF complex, we quantified CUX1 and SMARCA4 occupancy at the significantly down ATAC sites. Compared to CUX1-independent sites, occupancy of both CUX1 and SMARCA4 are significantly higher at significantly down ATAC sites (**Figure 4.5D**). Examples for significant loss of chromatin accessibility following CUX1 knockout are shown at NCOA4, which promotes erythropoiesis by regulating ferritin turnover,[313] and ADTRP, which regulates myelopoiesis and definitive hematopoiesis (**Figure 4.5E**).[314] Taken together, in human HSPCs, CUX1 is directly involved in maintaining chromatin accessibility at

enhancers associated with SMARCA4 occupancy and targeting genes regulating hematopoiesis.

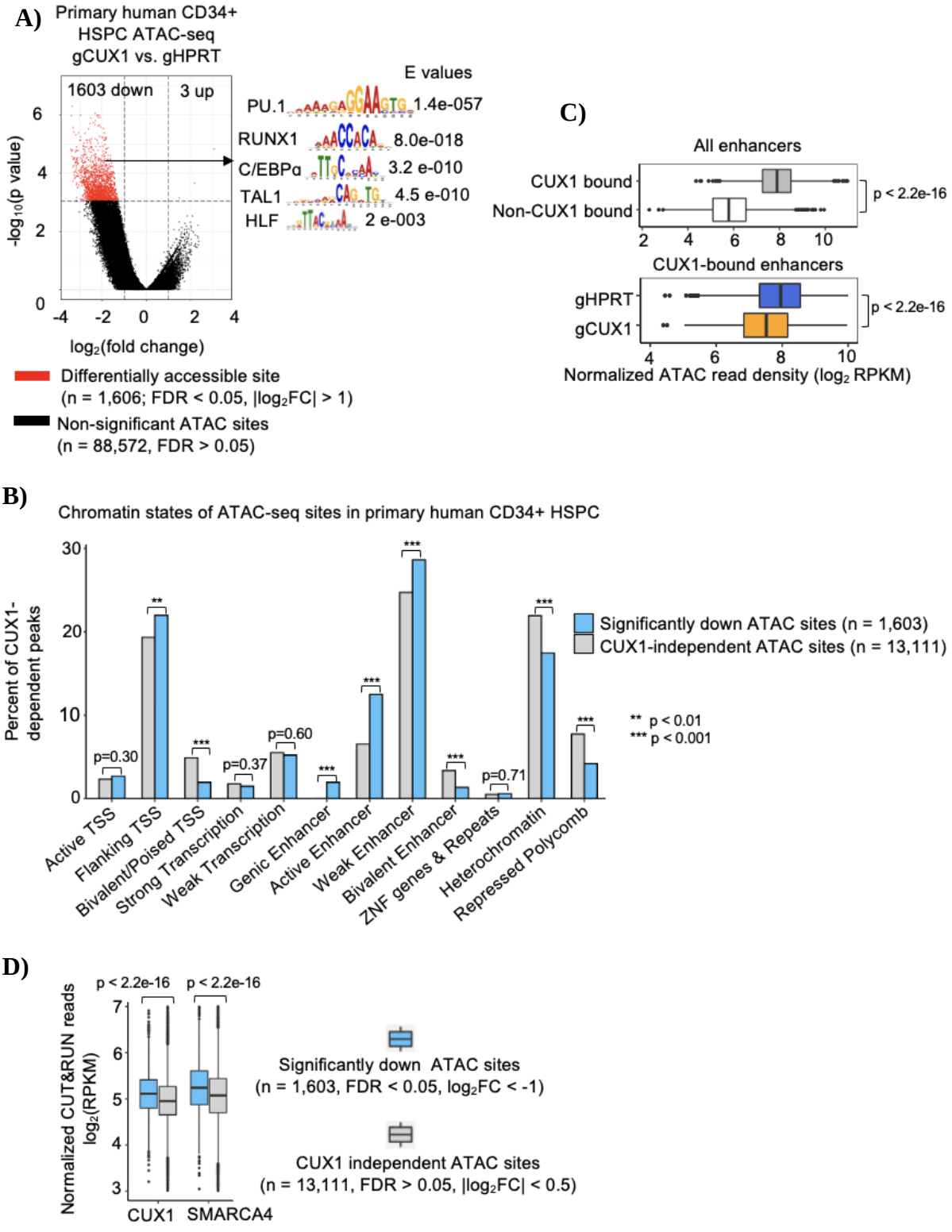


Figure 4.5: In human HSPCs, CUX1 and SMARCA4 maintain chromatin accessibility at enhancers associated with hematopoietic differentiation. **A)** Volcano plot of ATAC-seq changes in gCUX1 and gHPRT CD34+ HSPCs (n=2 biological replicates). Significance calculated by csaw.[232] Top motifs for the down sites are shown.

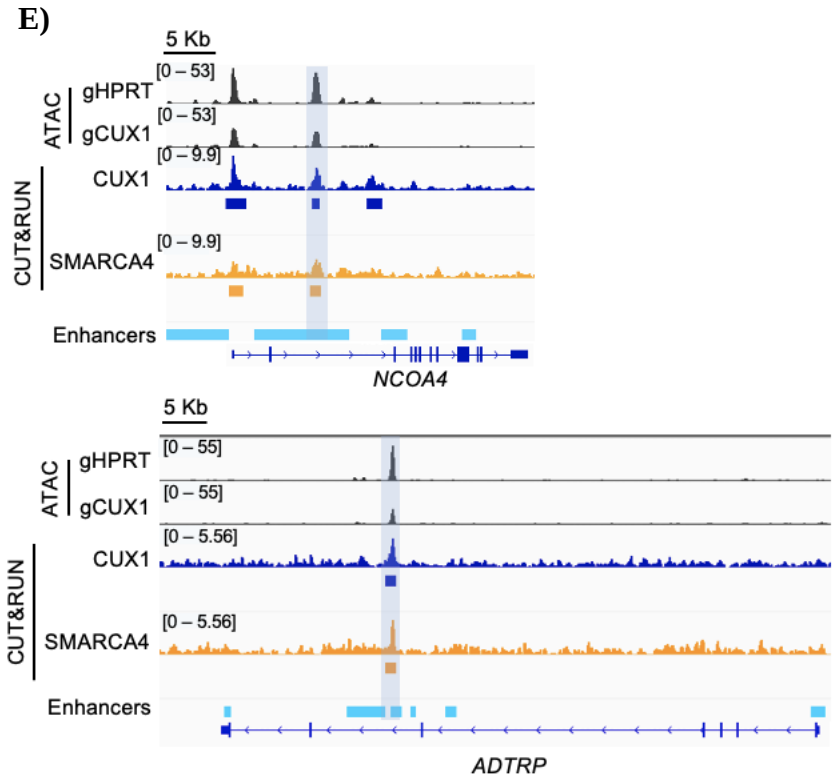


Figure 4.5 continued B) Distribution of chromHMM chromatin state of the peaks whose accessibility are significantly downregulated after CUX1 loss (n=1,603, blue, FDR<0.05, $\log_2$FC<-1) and the ATAC sites whose accessibility are not dependent on CUX1 (n=13,111, grey, FDR>0.1, |$\log_2$FC|<0.5). Significance is calculated using hypergeometric test. **C)** Normalized ATAC reads at genome-wide CUX1-bound enhancers (n=3,902) and a randomly sampled, size-matched list of enhancers not bound by CUX1 (top). Normalized ATAC reads at CUX1-bound enhancers (n=3,902) comparing the control gHPRT and gCUX1 conditions (bottom). **D)** Normalized CUT&RUN reads of CUX1 and SMARCA4 in CD34+ HSPC at down vs. CUX1-independent ATAC sites. Significance for C) and D) is by two-sided Wilcoxon rank-sum test. **E)** IGV analysis of normalized ATAC-seq signal (RPKM) tracks of gHPRT and gCUX1 cells at hematopoiesis-regulating genes NCOA4 and ADTRP. Normalized CUX1 and SMARCA4 binding signal (RPKM) from CUT&RUN experiment are shown along with CD34+ HSPC chromHMM enhancer annotations. (NCO4 lost site log2FC=-1.16, FDR=0.073; ADTRP lost site log2FC=-1.49, FDR=0.027)

4.2.4 CUX1 genomic targets are linked with genome architecture and in vivo lineage potential.

Previous studies reported that CUX1 binding is highly predictive of enhancer-promoter interactions.[169,289] As we observed a substantial proportion of CUX1 binding at promoter-proximal regions in human CD34+ cells (**Figure 4.4C**), we tested if CUX1 binding at these promoters influences accessibility at enhancers looped to those promoters. We intersected 2,684 looping DNA contact points, obtained from Hi-C analysis of human HSPCs,[237] with CUX1 CUT&RUN data and identified n=272 DNA loops that contain distal elements in contact with CUX1-bound promoters. Integrating these sites with our ATAC-seq data revealed two findings. First, distal elements in contact with CUX1-bound promoters had overall increased DNA accessibility as compared to non-CUX1-bound counterparts (**Figure 4.6A**). Second, distal elements in contact with CUX1-bound promoters trend towards decreased accessibility after CUX1 loss (p=0.085), while there is no change in accessibility for loops not in contact with CUX1-bound promoters (p=0.89) (**Figure 4.6A**). An example genome snapshot of CUX1 promoting accessibility of enhancers looped to CUX1-bound promoters is shown at cell surface protein tyrosine kinase KIT, which regulates stem cell self-renewal (**Figure 4.6B**).[315] In summary, CUX1 binding to promoters is associated with increased accessibility of looped enhancers.

Heretofore, our data suggest that CUX1 with SMARCA4 promotes accessibility for recruitment of TFs that drive differentiation (**Figure 4.5A**). To explore the transcriptional consequences of CUX1 loss, we integrated the ATAC-seq with RNA-seq from CD34+ HSPCs with 98 genes upregulated (FDR<0.1, log2FC>0.75) and 334 genes downregulated (FDR<0.1, log2FC<-0.75) after CUX1 knockdown.[79] In total, 406/432 of the differentially expressed gene (DEGs) contain significantly decreased ATAC-seq sites. Of these 406 genes, 317 have decreased while only 89 have increased expression (**Figure 4.6C**). The proportion

of DEGs with simultaneously decreased RNA expression and DNA accessibility is significantly higher than random ($p<2.2e-16$, chi-squared test). This finding links CUX1-dependent increased DNA accessibility with increased target gene expression, as expected. Notably, both CUX1 and SMARCA4 occupancies are higher at these 406 genes than in the background control, demonstrating a positive correlation between the presence of CUX1 and BAF in chromatin accessibility and RNA expression (**Figure 4.6D**). While gene ontology (GO) enrichment analysis revealed no significantly enriched GO terms for the 89 genes with increased RNA levels, those genes that decreased were enriched for genes involved in lineage potential and transcriptional priming (**Figure 4.6C**). Therefore, our data indicate that the chromatin accessibility-promoting role of CUX1 in human HSPC is coupled to transcriptional changes in lineage potential.

Lineage-determining TFs bind enhancers to drive cell-type specific gene expression and terminal differentiation.[109,316] Based on the evidence that CUX1 regulates HSPC cell fate in driving erythroid, myeloid and lymphoid fate decisions,[79] we hypothesized that CUX1 promotes accessibility at cell-type specific enhancers. We obtained a list of enhancer annotations specific for each human hematopoietic cell type from the Integrative and Discriminative Epigenome Annotation System (IDEAS) database of the VISION project.[233] We then quantified the change in accessibility after CUX1 editing at these enhancers (**Figure 4.6E**). Loss of CUX1 induced a significantly larger drop in accessibility at cell-type specific enhancers for all hematopoietic lineages, compared to the control, which is a randomly sampled (n=10,000) set of enhancers that did not appear in any cell-type specific enhancer lists ($p<2.2e-16$). These data suggest that CUX1 preferentially unmasks DNA at lineage-specific enhancers to facilitate hematopoietic maturation.

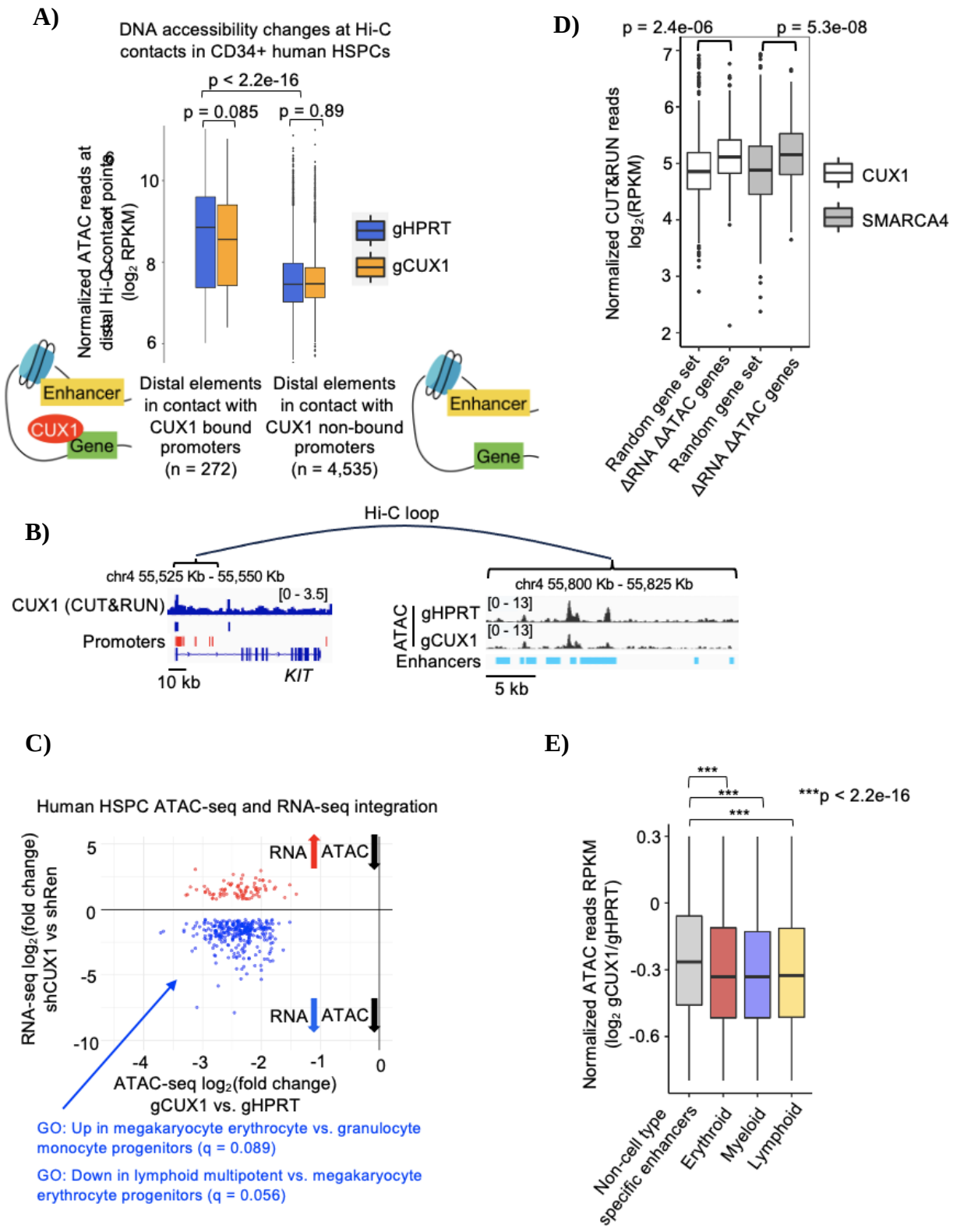


Figure 4.6. CUX1 genomic targets are linked with genome architecture and *in vivo* lineage potential **A)** ATAC-seq accessibility for gHPRT and gCUX1 CD34+ HSPCs at distal 3D chromatin contact points looped to CUX1-bound promoters from published CD34+ HSPC Hi-C data.[237] **B)** IGV snapshot of CUX1 binding at the promoter of KIT and the reduced accessibility of multiple enhancers looped to the promoter. Enhancer and promotor

Figure 4.6 continued annotations are from Roadmap Epigenomics.[236] **C)** Integration of CD34+ HSPC ATAC-seq and RNA-seq (n=2 biological replicates).[79] Scatterplot shows the RNA log2FC vs. ATAC-seq log2FC for 406 DEGs (FDR<0.1, |log2FC|>0.75). Enriched GO terms related to HSPC lineage commitment are shown.[317] **D)**A quantitative comparison of CD34+ HSPC normalized CUT&RUN reads of CUX1 (left) and SMARCA4 (right) at the sites with simultaneously significant changes in RNA expression and chromatin accessibility from Figure 4C (n=406), vs. the control, which are size-matched regions associated with randomly sampled genes (n=406) from csaw results. **E)** Log2FC of ATAC-seq signal comparing CD34+ HSPC gCUX1 vs. gHPRT cells at the hematopoietic cell-type specific enhancers from the VISION database[318] (9,657 myeloid enhancers, 11,653 erythroid enhancers, 15,323 lymphoid enhancers), and 10,000 randomly sampled non-cell type specific enhancers.

Development in single cell sequencing technology including scRNA-seq provided us with a rich toolbox to discover the gene expression heterogeneity in early HSCs. Through clonal lineage tracing, we can for the first time link the transcriptome of ancestral HSCs to their progeny cells.[319] This enabled us to interrogate how well genes regulated by CUX1 in HSCs can predict cell fates. To examine the role of CUX1 gene regulation in cell fate decisions *in vivo*, we turned to a clonal lineage-tracing dataset coupling murine HSPC single-cell transcriptomic state to progeny cell fates.[238] To test the hypothesis that CUX1 target gene expression is predictive of lineage determination, we used logistic regression and deep neural network classifiers as described by Weinreb et al 2020.[238] Comparable to their studies, a randomly sampled group of genes (n=1,000) and a curated list of mouse TFs (negative controls) returned less than 50% prediction performance measured by F1 score, whereas the top 1,000 most variable genes (positive control) returned F1 scores of ~61-63%, validating our machine learning models (**Figure 4.7A**). While all CUX1-bound genes we identified in HSPCs (n=6,758) could not predict cell fate well (F1<45%), CUX1-bound genes with differential expression after CUX1 knockdown (n=923) improved accuracy to 56-58%. Notably, this performance is similar to the published equivalent gene sets from known HSPC

fate-specifying pioneer factors PU.1 (45-53%) and RUNX1 (51-57%) (**Figure 4.7A**).[242–245] This analysis suggests that CUX1 regulated genes are predictive of HSPC cell fate in vivo.

Since CUX1 disrupts HSC homeostasis and controls the severity of myeloid diseases in a dosage dependent manner,[79] an interesting question is whether the gene expression changes that drive these phenotypes manifest early in hematopoietic hierarchy. To this end, we generated a shRNA knock-down mouse model under bone-marrow transplant setting, where *Cux1* expression is reduced to medium (54% +/-17%) and low (12% +/-9%) levels respectively. We isolated the Kit+Lin- HSPC population through flow cytometry. scRNA seq identified distinct progenitor clusters under different Cux1 dosages (**Figure 4.7B**). Differential abundance analysis revealed that the erythroid progenitors (collectively labeled as MEP, or megakaryocytes and erythrocyte progenitors) are significantly less abundant in both $Cux1^{mid}$ and $Cux1^{low}$ conditions (**Figure 4.7C**). The labelled MEP population consist of both erythroid progenitors and megakaryocyte progenitors expressing platelet factor 4 (*Pf4*). This observation is consistent with the fact that knocking down *Cux1* led to impaired erythroid differentiation and anemia in mice,[79] and implied that the effect of erythropoiesis blockage upon Cux1 loss is manifested in hematopoietic apex.

Next, I sought to investigate the transcriptomic effect of Cux1 dosages on the HSPC population. In the Renilla control sample, I binned all 9074 cells into three groups by the level of *Cux1* transcripts. The high *Cux1* group contains cells whose Cux1 expression fall into the top 5 percentile, and the middle group contains cells in the 80%-95% percentile of *Cux1* expression. The low group contains size-matched 1000 cells by random sampling from the bottom 80% of *Cux1* expression (**Figure 4.8A**). Correlation of expression of all genes in these three groups revealed that cells with lower *Cux1* expression have more correlated gene expression modules that contain lineage-specific markers, while the cells with the highest level of *Cux1* expression are devoid of such cell-type specific modules (**Figure 4.8A**). This

observation in the endogenous population implies that Cux1 level is positively correlated with the stem-like transcriptional program. To investigate how Cux1 dosage change affected the gene expression of each progenitor population, I pseudobulked the UMIs for each gene across all cells in each cluster, and aggregated them into a single count. Differential expression analysis revealed that $Cux1^{low}$ lead to most DEGs in neutrophil progenitor and common lymphoid progenitor (B cell biased) clusters, while $Cux1^{mid}$ lead to most DEGs in MEP clusters. This analysis showed that *Cux1* affects the HSPC progenitor gene expression in a lineage-biased and dosage dependent manner.

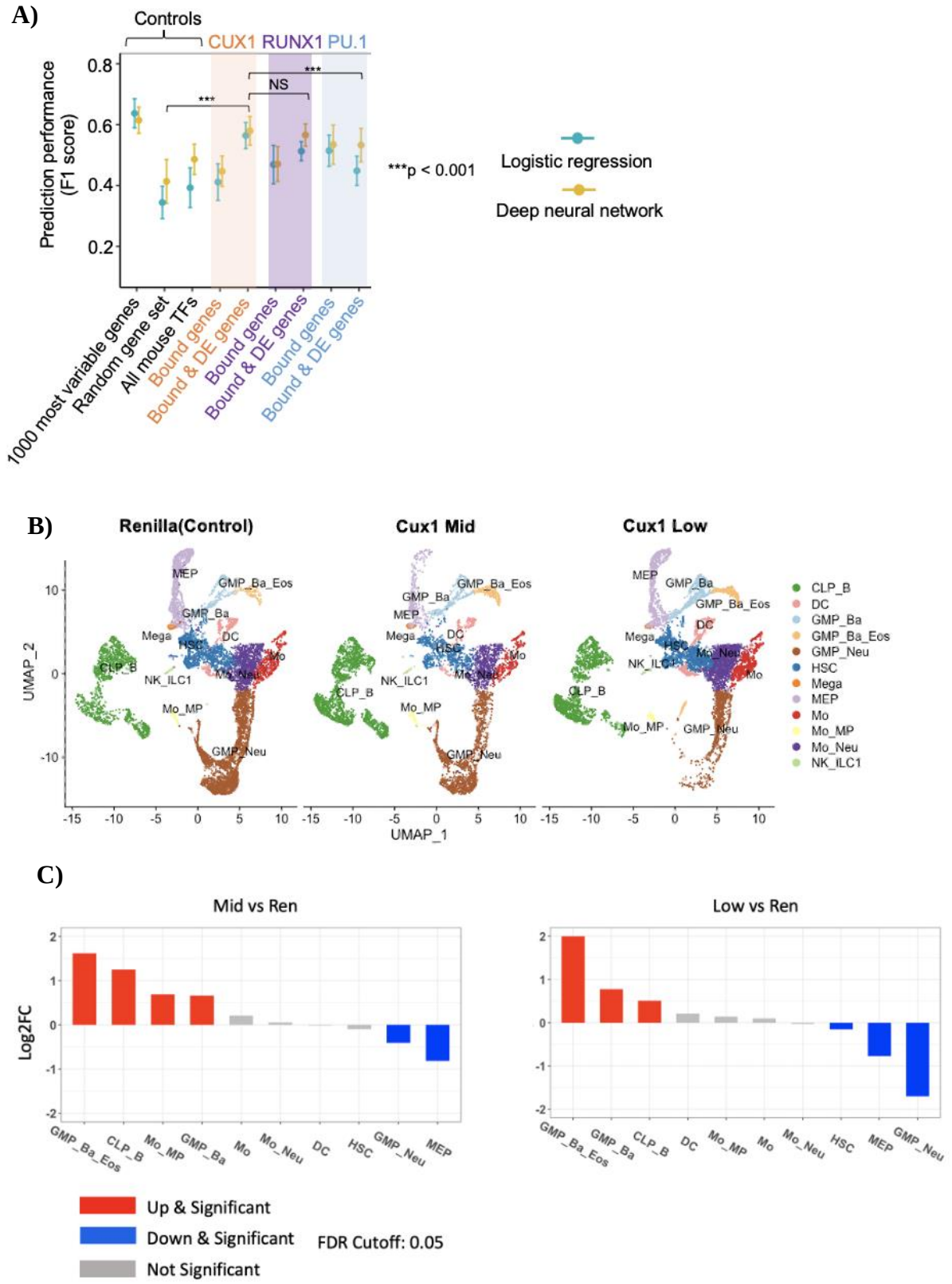


Figure 4.7. CUX1 genomic targets predicts HSPC cell fate and loss of CUX1 lead to lineage imbalance **A)** Performance score of cell fate prediction using the published murine HSPC scRNA-seq.63 From left to right: positive control is the top 2,000 genes with the highest cell-cell variation; negative controls are a randomly sampled gene set (n=1,000) and curated list

Figure 4.7 continued of mouse TFs (n=1,636);99 CUX1-bound genes from human CD34+ HSPC CUT&RUN (n=6,758); overlap of CUX1-bound and differentially-expressed upon CUX1 knockdown in CD34+ HSPC (n=923). Equivalent gene sets were tested for PU.1 and RUNX1 as benchmarks (n = 336 and 325). For all gene sets larger than 1,000, 50 bootstraps were performed to sample for 1,000 genes. Logistic regression and deep neural network were used to construct the classifier. Significance is calculated using two-sided Wilcoxon rank sum test. **B).** UMAP representation of the shRenilla control, $Cux1^{mid}$ and $Cux1^{low}$ populations in scRNA-seq. Each dot represents a single cell after quality control thresholds outlined in the method section. The top up-regulated DEGs were used to label the identity of each cluster. Clusters expressing the lineage markers of specific hematopoietic cell types are labelled accordingly. **C)**Differential abundance test on the abundance of each progenitor cluster's cell count comparing $Cux1^{mid}$ to Renilla control (left) and $Cux1^{low}$ to Renilla (right). Differential abundance model and significance test is calculated using EdgeR framework.[254] (FDR < 0.05)

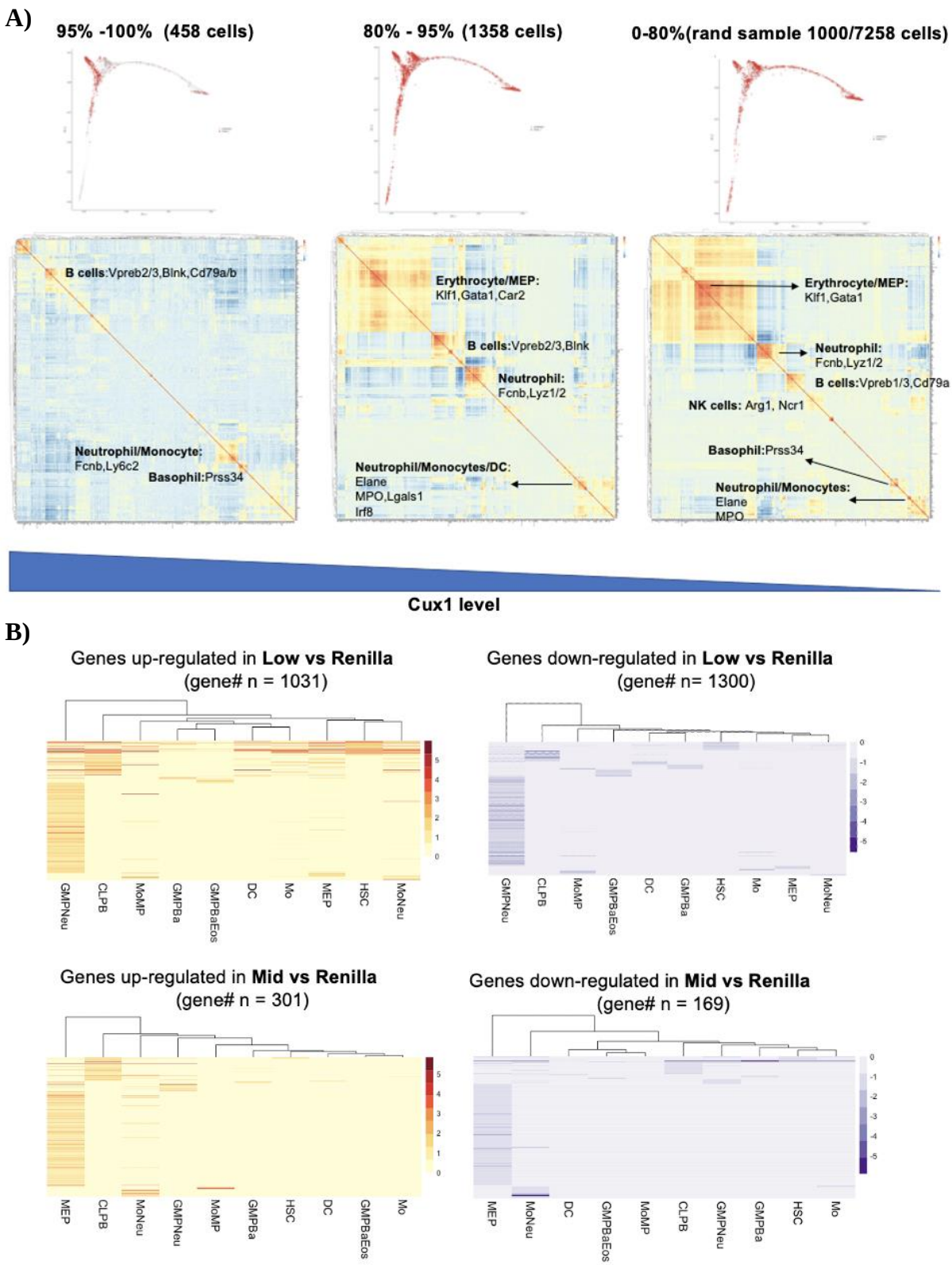


Figure 4.8. CUX1 regulates lineage-specific HSPC transcriptome in a dosage dependent manner **A)** Spearman correlation heatmap of all detected genes in Renilla sample. The three heatmaps show the correlation for the subpopulations that express high, medium and low levels of Cux1, respectively. For the low Cux1 bin because there are way more cells than the high and medium group, I randomly sampled 1000 cells in order to size match the sample

Figure 4.8 continued size in high and mid groups to ensure proper statistical power. **B)** Differential expression analysis heatmaps between progenitor populations from $Cux1^{Low}$ vs Renilla and $Cux1^{mid}$ vs Renilla. Up and down regulated DEGs are plotted in separate heatmaps. Value plotted are log_2 transformed fold change. Cut-off used are FDR < 0.1 and |log2FC| > 1.

4.3 DISCUSSION

Our study delineates a mechanism whereby CUX1 facilitates chromatin remodeling by recruiting the BAF complex to enhance DNA accessibility, a pivotal process in gene regulation (**Figure 4.9**). Nucleosomes, which typically obstruct transcription factors (TFs) and RNA polymerase from DNA access, are destabilized by CUX1, albeit without causing displacement.[170,171] This destabilization primes for further remodeling by the BAF complex, as observed in cellular contexts. This interaction is evident in a significant portion of CUX1 binding sites, demonstrating a "direct model" of CUX1-dependent SMARCA4 recruitment and increased DNA accessibility (**Figures 4.2B and 4.3F**). An alternative "indirect model" also exists, where CUX1 binding does not necessitate SMARCA4 recruitment, suggesting the involvement of other TFs like SP1 at these sites,[287,320] and raises questions about the mechanisms by which CUX1 influences chromatin accessibility indirectly (**Figures 4.2B and 4.3G**).

In the context of hematopoiesis, CUX1's role as a pioneer factor essential for epigenetic reprogramming of stem and progenitor cells is highlighted, marking it as a critical player in the early differentiation stages in primary human HSPCs.[321,322] This role suggests a broader function for CUX1 in regulating chromatin accessibility across various tissues and implies its general regulatory capacity over enhancer activation by lineage-specific TFs.

CUX1 haploinsufficiency is associated with developmental abnormalities and myeloid malignancies, indicating its essential role in genomic stability. The impact of CUX1 haploinsufficiency on BAF recruitment, alongside the possibility of altered BAF activity due

to CUX1 loss, warrants further investigation. This will be discussed in detail in the final discussion section.

The research underscores CUX1's indispensable role in hematopoiesis and its implications in hematopoietic disorders, suggesting that targeting BAF complex activity could provide therapeutic benefits in CUX1-deficient myeloid neoplasms. This comprehensive understanding of CUX1 and BAF's roles in chromatin remodeling and gene regulation presents a foundation for further investigation into their biological functions and therapeutic potential.

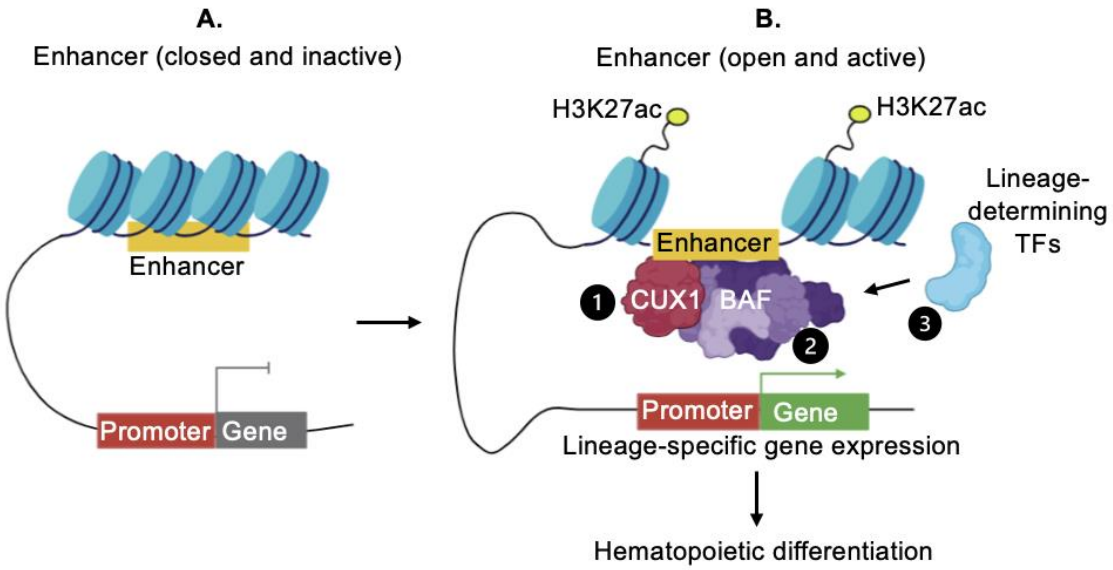


Figure 4.9: Working model schematic of CUX1 pioneer function. In human HSPC, CUX1 acts as a pioneer factor to closed enhancers and recruits the BAF chromatin remodelling complex to promote DNA accessibility. Subsequently CUX1 recruits or collaborates with other hematopoietic TFs to promote lineage specific gene expression programs that ensure proper hematopoietic differentiation. The graph was created using biorender (https://www.biorender.com/).

CHAPTER 5: CUX1 SERVES AS A GATEKEEPER FOR GATA1-MEDIATED ERYTHROID DIFFERENTIATION

5.1 INTRODUCTION

Research at our lab showed that CUX1 promotes healthy erythropoiesis. *Cux1* haploinsufficient mice develop MDS with anemia and tri-lineage dysplasia.[79] Based on my work in the previous section, CUX1 act as a pioneer factor in hematopoietic apex to recruit chromatin remodeler and promote pro-erythroid transcriptional programs. At the same time, GATA1 is a known central regulator expressed in early erythroid progenitors essential for maintaining healthy erythropoiesis.[197] Furthermore, CUX1 and GATA1 frequently co-localize with each other. CUX1 binding sites are significantly enriched with GATA motif. ChIP-seq and CUT&RUN confirmed that there are significant overlaps between genome-wide CUX1 and GATA1 binding sites in both K562 and primary human HSPCs ($p < 2.2E-16$, hypergeometric test, **Figure 5.1A**). Therefore, two exciting outstanding questions remain: 1) What are the exact mechanisms of CUX1 loss that impair erythropoiesis 2) What the interaction mechanism between CUX1 and GATA1 is in HSPC, and how does their interaction mechanism affect HSPC cell fate determination. Do they collaborate with or recruit each other, or do they antagonize each other?

5.2 RESULTS

As part the same manuscript, Angela Stoddart and others at our lab has performed a series of studies investigating the mechanism of how CUX1 loss impair erythropoiesis, I'm going to summarize them briefly in the first two sections:

5.2.1 CUX1-knockdown uncouples erythroblast cell division from differentiation.

During erythropoiesis, cell division is tightly coupled with differentiation with each daughter cell functionally different than the mother cell.[172] The cell surface marker CD44

with forward side scatter (FSC) can be used to show that proerythroblasts sequentially generate basophilic, polychromatic and orthochromatic erythroblasts following a doubling 1:2:4:8 ratio.[323] Using the same shRNA transgenic mice model using in An et. al 2018[79], our lab found that while in the shRenilla control mice, the expected doubling ratio of each successive population is observed (1.0, 2.7, 5.8, 9.4), in $Cux1^{low}$ low, this ratio is disrupted (1.0, 3.3, 9.7, 10.1), suggesting CUX1 loss uncoupled erythroid differentiation with proliferation. Furthermore, bone marrow cells from the mice are sorted using surface markers into four sequential erythroblast populations RI (CD71hi/Ter119med), RII (CD71hi/Ter119hi), RIII (CD71med/Ter119hi), and RIV(CD71-/Ter119hi). In $CUX1^{low}$ mice, there is a decrease of RII (basophilic) population (38% down to 22% in terms of total population), but an increase of the RIII (polychromatic) population. However, there is no to little change in the RIV population. This observation suggesting that loss of CUX1 disrupts the erythroid differentiation trajectory.

5.2.2 CUX1 knockdown leads to vast opening of erythroblast chromatin and transcriptional deregulation of GATA1-target genes.

Erythroid progenitors undergoing terminal erythropoiesis must undergo vast chromatin condensation in preparation for enucleation.[324] Differential accessibility analysis of ATAC-seq on the RII (CD71+Ter119+) population isolated from Renilla control and $CUX1^{low}$ showed that knocking down CUX1 leads to vast chromatin opening (7426 significantly up peaks comparing to 96 significantly down peaks, called by csaw.[232] Gene ontology enrichment analysis found that the genes associated with significantly up peaks encodes proteins with negative regulation of erythroid differentiation, and acetyltransferase activity, such as Ep300 (P300/KAT3B) and Crebbp (CBP/KAT3A). Western blot confirmed

that acetylation of several histones including H3K18, H3K23, H3K27 and H4K12 were increased >2 fold in erythroblasts following CUX1 KD. These results suggest that CUX1 loss might impair healthy erythropoiesis through reversing the chromatin condensation process of terminal erythropoiesis. Histone acetylation may play a critical role in driving up an open chromatin state in erythroblasts after CUX1 KD.

Although CUX1 KD did not significantly alter the *Gata1* gene expression, gene set enrichment analysis showed that CUX1 is required for GATA1 activated genes. GATA1 was first identified as a protein with binding specificity to the β-globin 3' enhancer.[325] Fetal hemoglobin genes, not normally expressed in adult erythropoiesis, were among the top upregulated genes after CUX1 KD, suggesting a major de-regulation of GATA1-mediated hemoglobin transcription. Together, these data implicate a functional interaction of CUX1 with GATA1 in driving terminal erythroid differentiation.

5.2.3 CUX1 binding dynamically shifts to GATA1-bound erythroid specific enhancers.

To explore the role of CUX1 in human erythropoiesis, we used a well-defined model where human CD34+ HSPCs are expanded and differentiated *ex vivo* into stage-matched populations of erythroid progenitor cells (**Figure 5.1B**). Using published RNA-seq data we illustrate that the *PU.1, TAL1* and *GATA1* transcription factors are dynamically expressed, with *PU.1* decreasing and *TAL1* and *GATA1* increasing as stem cells give rise to erythrocytes, consistent with their roles in lineage specification.[326] *CUX1* expression also changes; it is highest in HSPCs (day 0), decreases as cells commit to the erythroid lineage and then gradually increases (**Figure 5.1C**).

We next assessed CUX1 binding, using CUT&RUN, at days 0 and 7 which correspond to undifferentiated HSPCs and basophilic erythroblasts. A differential motif enrichment analysis of CUX1 bound sites at day 0 versus day 7 suggests that in human stem

cells CUX1 binding is shared with several ETS TF family members (ETS1, ERG, FL1, PU.1) and shifts to the erythroid transcription factors, GATA1 and TAL1, as cells commit to the erythroid lineage (**Figure 5.1D**). Leveraging ChIP-Seq occupancy of the lineage-regulated TFs, PU.1, GATA1 and TAL1,[298] we show that prior to erythroid commitment, CUX1 binding overlaps more with PU.1 than GATA1 and TAL1 (**Figure 5.1E**). At day 7, when cells become erythroblasts, 50% (1850/3695) and 47% (1740/3695) of CUX1 binding sites overlap with GATA1 and TAL1 compared to only 5% (314/6758) and 4% (244/6758) at day 0. These observations highlighted the dynamic shift in CUX1 binding partners as cells commit to the erythroid lineage (**Figure 5.1E**).

Using histone modifications H3K4me1 and H3K27Ac J. Huang et al 2016. previously identified HSPC-specific enhancers (enriched for FLI1 and PU.1) and erythroblast-specific enhancers (enriched for GATA1 and TAL1).[203] We found that ~10% of CUX1 binding sites overlap with HSPC-specific enhancers prior to lineage commitment and ~35% of CUX1 binding sites overlap with erythroblast-specific enhancers after cells differentiate into erythroblasts (**Figure 5.1F**). An example genome snapshot illustrating the gradual increase of CUX1 enhancer binding was shown at the enhancer locus of *SLC25A37,* which encodes an essential iron importer for mitochondrial heme biosynthesis in erythroblast.[327] (**Figure 5.1G**) Together this suggests that CUX1 binding dynamically shifts to GATA1-bound erythroid specific enhancers in human erythroblast.

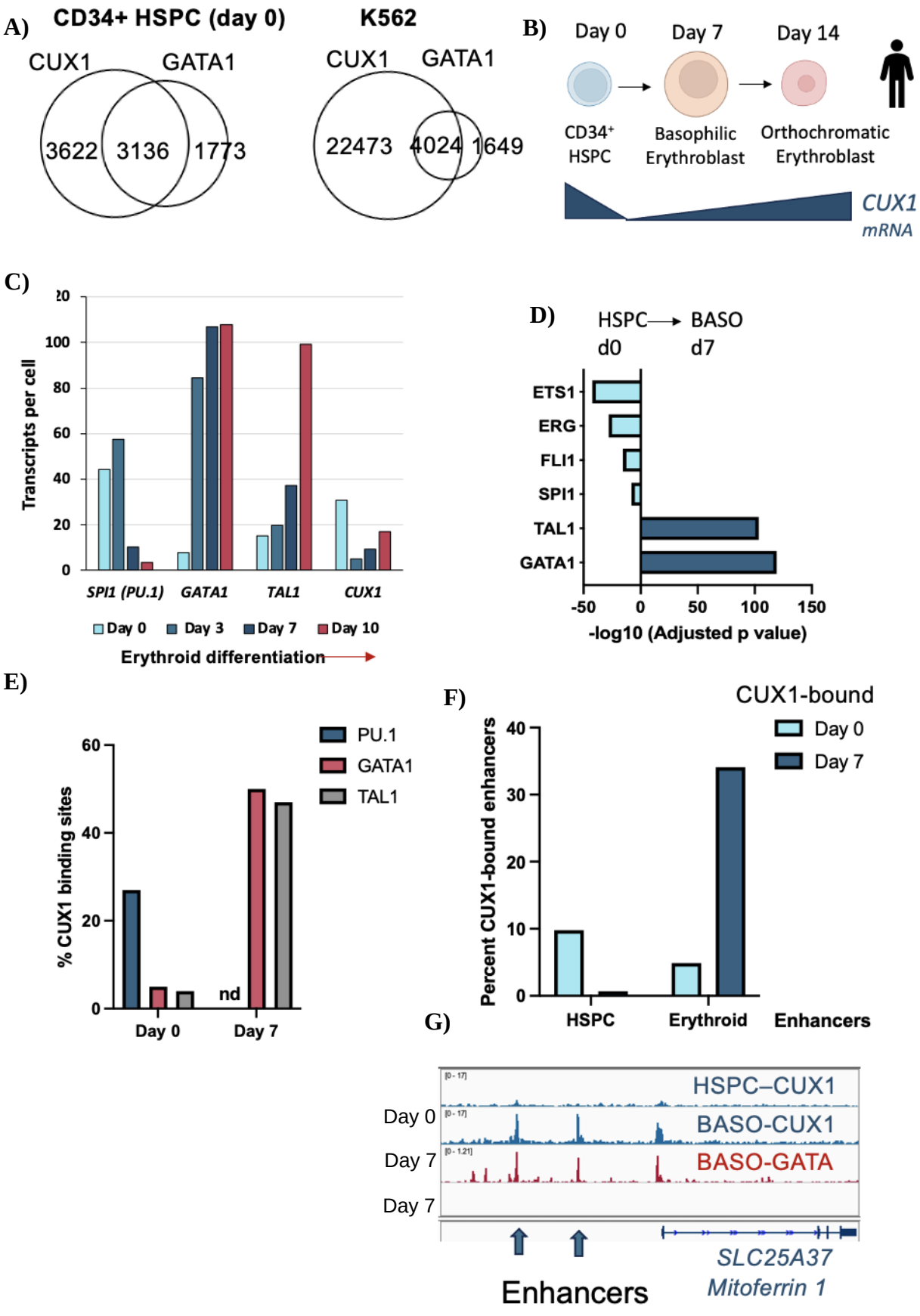


Figure 5.1 CUX1 binding dynamically shifts to GATA1-bound erythroid specific enhancers

Figure 5.1 continued A) Overlap of CUT&RUN peaks of CUX1 and GATA1 in primary CD34+ human HSPCs (left). Right panel shows the overlap between CUT&RUN GATA1 peaks and ChIP-seq called CUX1 peaks in K562 cell. **B)** Schematic of the erythrocyte differentiation used for RNA-seq and CUT&RUN. Cells were harvested at day 0 (HSPC), day 7 (basophilic erythroblasts) and day 14 (orthochromatic erythroblast). **C)** Gene expression (Transcripts per cell) of hematopoietic TF PU.1, TAL1, GATA1 and CUX1 at day 0, 3, 7 and 10.(PMID: 21845190) **D)** Top significantly enriched motifs from differential motif analysis of CUX1 binding sites at day 0 (HSPC, left) and day 7 (basophilic erythroblast, right) of erythroid differentiation. **E)** Percentage of CUX1 binding sites at day 0 and day 7 of erythroid differentiation that are co-occupied by PU.1, GATA1 and TAL1. **F)** Percentage of CUX1 binding sites that overlap with HSPC-specific and proerythroblast specific enhancers.[203] **G)** Example IGV genome snapshot at *SLC25A37* showing the increased CUX1 binding from day 0 to day 7 of erythroid differentiation. GATA1 binding at day 7 was also showed.

5.2.4 CUX1 loss leads to misdirected GATA1 binding.

To investigate how CUX1 loss impacts genome-wide GATA1 binding, we performed GATA1 CUT&RUN with the gHPRT and gCUX1 transfected K562 cells (two biological replicates each). Differential binding analysis showed that knocking out CUX1 leads to significantly more GATA1 binding sites with increased (n = 7522) than decreased (n = 87) occupancy (FDR < 0.05, $|Log_2FC| > 1$) (**Figure 5.2A**).[232] Intersecting the 7522 significantly upregulated GATA1 peaks with the endogenous GATA1 binding sites called from our lab and ENCODE showed that the vast majority of the significant up GATA1 peaks are not in regions where GATA1 usually bind and are thus "*de novo*" (**Figure 5.2B**).[298] In the subsequent text, these 7522 significantly up GATA1 sites will be referred as "*de novo* GATA1" sites. Genome-wide correlation analysis showed that higher endogenous CUX1 binding intensity are correlated with higher degree of increase of GATA1 binding upon CUX1 loss, implying CUX1 dosage is important in gatekeeping GATA1 binding (**Figure 5.2C**). Further characterization of the *de novo* GATA1 sites showed that they are mostly located at enhancer regions (**Figure 5.2D**). In addition, the K562 genomic loci that showed a significant increase in chromatin accessibility after knocking out CUX1 are enriched for the

GATA motif and have a significant higher magnitude of GATA1 occupancy increase compared to other genomic loci (**Figure 5.2F**), implying that the CUX1-dependent DNA accessibility might prevent abnormal GATA1 binding in the endogenous condition.

Functionally, GATA1 could either serve as a repressor or activator for gene expression in a context dependent manner.[187] To investigate the transcriptional consequences of the *de novo* GATA1 binding, we integrated the GATA1 CUT&RUN data with RNA-seq data in K562 cells with gHPRT and gCUX1. Upon knocking out CUX1, genes who have a simultaneous increase in RNA expression and GATA1 occupancy increase are enriched for mitosis and cellular proliferation pathways (**Figure 5.2E**), while the genes who have GATA1 occupancy increase but RNA expression decrease are enriched for regulation of erythroid differentiation pathways. This analysis is consistent with the fact that loss of CUX1 leads to increased HSPC proliferation and impaired erythropoiesis,[79] thus providing a potential mechanistic explanation that the abnormal *de novo* GATA1 binding might be a cause behind these phenotypes.

To determine if the *de novo* GATA1 binding are direct or indirect consequence of CUX1 occupancy, we divided the non-significant GATA1 sites and *de novo* GATA1 sites into those bound and not bound by CUX1 (**Figure 5.3A**). Both the CUX1 bound and non-CUX1 bound *de novo* GATA1 sites have significant increased GATA1 binding. This indicates that under endogenous condition, CUX1 might gatekeep and prevent GATA1 binding at these sites through both physically blocking GATA from accessing and through unknown indirect mechanisms. For the sites where CUX1 directly shield GATA1 from binding (n = 2,914), they are more distributed at distal enhancer elements are enriched for hematopoietic TF motifs including *SPI* (encoding PU.1), *FLI1* and *RUNX1* (**Figure 5.3B, C**). This observation implies that these sites might be essential hotspots for hematopoietic differentiation regulation and warrants further investigation.

The ChIP-seq signal of the active enhancer histone marks H3K4me1 and H3K27ac increased significantly at both the CUX1 bound and non-bound de novo GATA1 sites after CUX1 KO (**Figure 5.3D**). This could result in a more permissive active enhancer chromatin environment for GATA1 to bind, and thus be a mechanistic explanation underlying the *de novo* GATA1 binding. However, the exact molecular mechanism of how loss of CUX1 lead to increased H3K4me1 and H3K27ac deposition remains unknown and worth further investigation.

In summary, these data show that CUX1 serves as a gatekeeper to prevent promiscuous GATA1 binding under normal conditions, possibly through both directly blocking GATA1 binding and indirectly mechanisms.

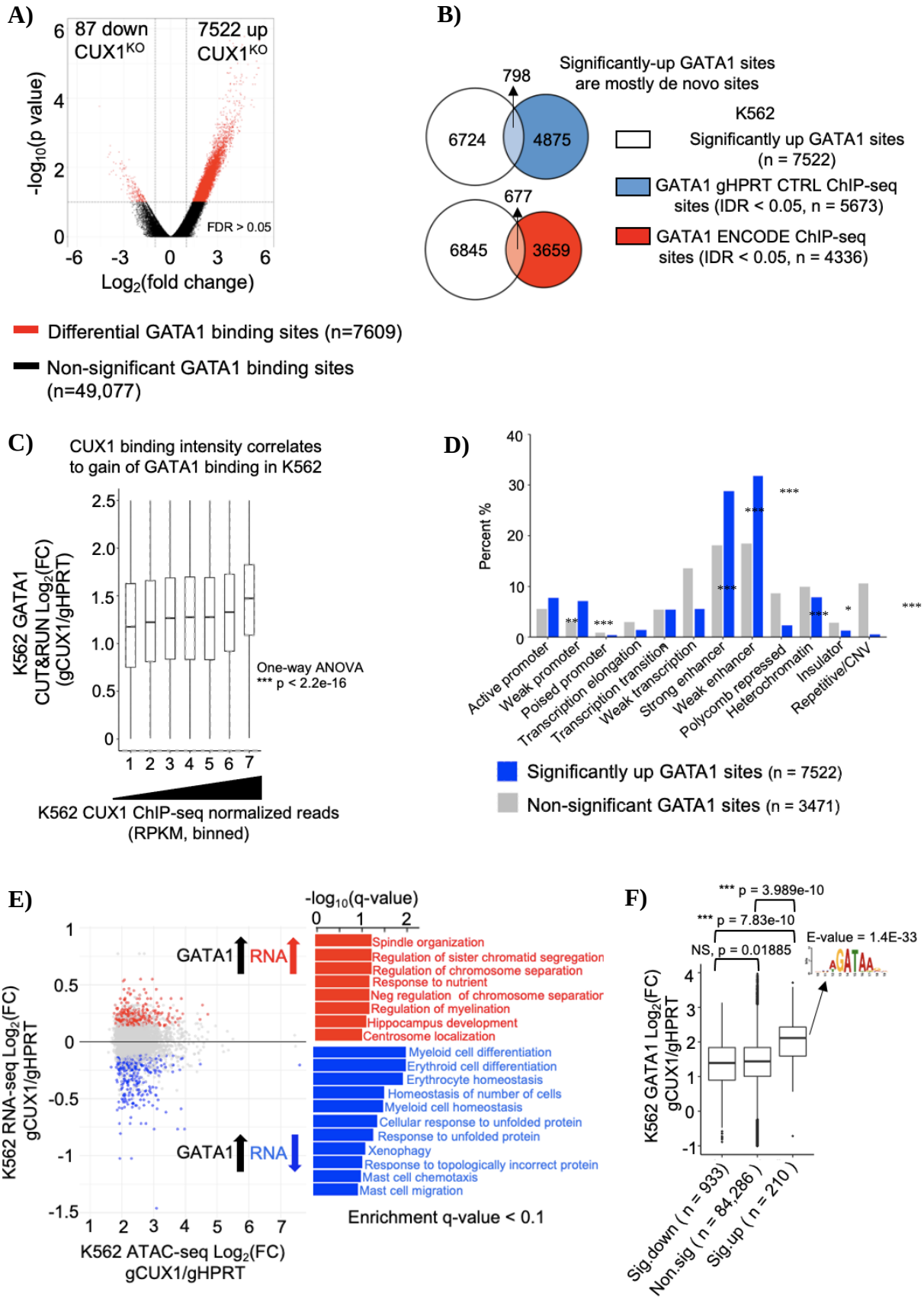


Figure 5.2. CUX1 loss leads to misdirected GATA1 binding. **A)** Volcano plot of the GATA1 CUT&RUN changes in gHPRT and gCUX1 (n = 2 biological replicates). Significance calculated by csaw.[232]

Figure 5.2 continued B) Overlap of number of de novo GATA1 sites (n = 7522) and wild type GATA1 binding sites called by our own CUT&RUN (top) and by ChIP-seq from ENCODE database (bottom) in K562. Wild type GATA1 peaks are called by MACS2 followed by IDR protocol (IDR threshold < 0.1). **C)** Genome-wide correlation of normalized CUX1 binding reads and log_2 fold change of normalized GATA1 reads in gCUX1 vs gHPRT conditions. Normalized CUX1 CUT&RUN reads were divided into 10 equidistant bins to represent the regions with lowest to highest CUX1 binding signal. The significance of the consistent upward trend is calculated using one-way ANOVA test. **D)** Distribution of chromHMM chromatin state of *de novo* GATA1 sites (n=7,522, blue, FDR<0.05, $log_2FC> 1$) and the GATA1 sites whose occupancy is not dependent on CUX1 (n=3,471, grey, FDR>0.1, $|log_2FC|<0.5$). Significance is calculated using hypergeometric test. **E)** Integration of K562 GATA1 CUT&RUN and RNA-seq (n=2 biological replicates). Scatterplot shows the RNA log_2FC vs. GATA1 CUT&RUN log_2FC for the 7,522 *de novo* GATA1 sites (FDR<0.1, log2FC>1). The sites who correspond to up and down differentially expressed genes are highlighted in red and blue, respectively (RNA-seq FDR < 0.1, $Log_2FC > 0$ and < 0). **F)** Log_2 fold change of GATA1 CUT&RUN signal after CUX1 KO for K562 genomic sites with significant down, non-significant and significantly up regulated chromatin accessibility.

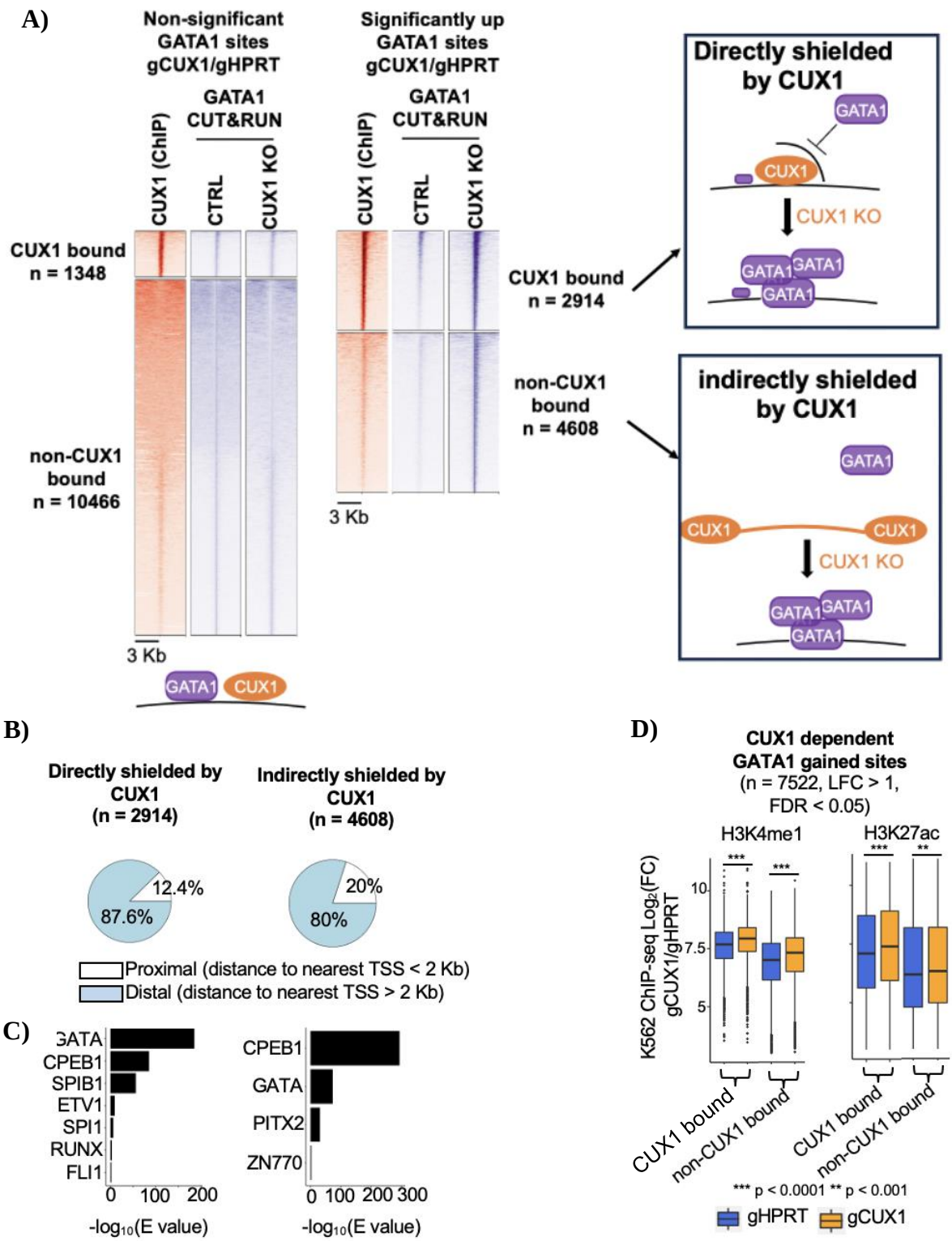


Figure 5.3: CUX1 gatekeeps GATA1 binding through direct and indirect mechanisms **A)** Heatmap showing the side-by-side K562 normalized CUX1 ChIP-seq and GATA1 CUT&RUN gHPRT & gCUX1 reads (RPKM) at the CUX1-independent non-significant GATA1 sites (FDR > 0.1, $|\log_2 FC| < 0.3$) and the de novo GATA1 sites (FDR < 0.05, $\log_2 FC > 1$). The peaks are divided by whether they overlap with CUX1 peaks (CUX1-bound) or not (non-CUX1 bound). **B)** Distribution of the distance to transcription start site

Figure 5.3 continued (proximal vs distal, threshold is 2 kB from TSS) for the de novo GATA1 sites bound vs not bound by CUX1, and **C)** the significantly enriched motifs at these two sets of sites. **D)** Normalized K562 ChIP-seq reads (RPKM) of the histone mark H3K4me1 and H3K27ac at the *de novo* GATA1 sites bound and not bound by CUX1, comparing the gHPRT vs gCUX1. Significance calculated by two-sided Wilcoxon rank sum test.

5.3 DISCUSSION

A complex network of TFs, cytokines and hormones orchestrate erythropoiesis.[177,178] Central to this process is the transcription factor GATA1. We had previously found that CUX1 deficiency impairs erythropoiesis at the expanse of myeloid expansion and can lead to anemia in mice as they age.[79] The mechanism whereby decreased CUX1 expression results in impaired production of red blood cells was not understood. My study provided a previously unknown interaction mechanism between two central regulators of hematopoiesis, CUX1 and GATA1.

Our research showed that CUX1 loss uncoupled the erythroblast differentiation and proliferation process and disrupted differentiation at the intermediate RII stage. Against the chromatin condensation normally required for erythrocyte reticulation and maturation, loss of CUX1 resulted in a vast chromatin opening in the primary mice RII cells and increased active histone acetylation. Transcriptionally, GATA1-regulated gene expression programs require normal dosage of CUX1 expression. This implies that CUX1 and GATA1 jointly regulate erythroid specific gene expressions that are essential for healthy erythropoiesis. Furthermore, integration analysis in human erythroblasts showed CUX1 binding undergoes a dynamic shift from mostly HSPC specific enhancers to GATA1-bound erythroid specific enhancers along the erythroid differentiation trajectory, which provided further regulatory roles of CUX1 in erythropoiesis.

Given the disruption in terminal erythroid differentiation after CUX1 loss, we originally hypothesized that GATA1 may not be binding its target genes in CUX1-deficient cells. However, we unexpectedly found that GATA1 binding was increased in the absence of normal CUX1 expression leading us to propose that CUX1 acts as a gatekeeper for GATA1 binding. The level of GATA1, examined by western blotting and intracellular staining, was not elevated in CUX1-deficient cell lines and cannot explain the increased binding. Some studies have proposed that acetylation of GATA1 itself could alter its binding,[328,329] but we found no evidence for increased or decreased GATA1 acetylation after CUX1 KD.

Mammalian terminal erythropoiesis involves gradual chromatin condensation steps that are essential for differentiation. The reason for this is that chromatin and nuclear condensation is followed by an enucleation step enabling the generation of a physically flexible mature red blood cell stage. $Cux1^{low}$ erythroid cells are capable of undergoing enucleation *in vitro* and there is no evidence for nucleated RBCs in mice (data not shown). It is therefore likely that the vast chromatin opening by ATAC-seq and increased histone acetylation observed in $Cux1^{low}$ erythroblasts alters accessibility of DNA binding proteins without causing huge structural changes to the nucleus.

Our study proposed that multiple regulatory mechanisms might exist to ensure GATA1 bind to the correct subset of loci. We find that the loss of CUX1 expression can dramatically alter GATA1 binding (**Figure 5.2A**). In some instances, CUX1 may be physically blocking access of GATA1 (direct model) but in other instances it may be indirect, such as changes to DNA accessibility and histone modifications (**Figure 5.3A**). We found that the fold increase of GATA1 binding is significantly higher in the genomic loci where accessibility increased after CUX1 loss (**Figure 5.2F**). This observation implies that the *de novo* GATA1 sites have increased accessibility upon CUX1 loss. However, whether the increase in GATA1 binding precedes or follows the increase in DNA accessibility remains

unknown. Previous study shown that GATA1 level does not caused genome-wide DNA accessibility change in erythroid progenitors using the G1E-ER4 cell line system.[330] Therefore, it is likely that it is the loss of CUX1 promotes DNA accessibility at the *de novo* GATA1 sites, and therefore allows more GATA1 binding. Future mechanistic studies should validate the sequence of these events and provide mechanistic explanation on why loss of CUX1 leads to such prominent GATA1 occupancy increase. In addition, our analysis also showed that the *de novo* GATA1 sites have significantly higher active histone mark H3K27ac and enhancer histone mark H3K4me (**Figure 5.3D**). The increased active histone modification is in concordance with the increased DNA accessibility, and these changes might co-ordinately provide a permissive chromatin environment for additional GATA1 binding. Going forward, it will be important to elucidate exactly how CUX1 promote the increased active histone mark deposition. For example, does CUX1 cooperate with the COMPASS family histone methyltransferase to increase the H3K4me1, and collaborate with HAT to increase H3K27ac? In addition, there might be additional indirect mechanisms that CUX1 utilizes to gatekeeps for GATA1 binding. Does CUX recruit other TFs and co-repressors to prevent GATA1 binding from the *de novo* GATA1 sites? Lastly, although most of *de novo* GATA1 binding sites have no detectable GATA1 binding events in the WT condition, there are some loci that already have some levels of GATA1 binding, but loss of CUX1 just further increased the GATA1 binding intensity at these sites. Our ongoing analysis is trying to dissect the transcription consequence of these two distinct types of gained GATA1 sites through integration with RNA-seq analysis.

The “gatekeeper” hypothesis is not in contradiction to our observation that GATA motifs are enriched in CUX1 binding sites in both K562 and HSPCs, as there are sizable overlaps between CUX1 and GATA1 binding sites in both cell types. Since CUX1 and GATA1 both promote healthy erythropoiesis, it is likely that the mode of interaction for these

two TFs are multifaceted. CUX1 might collaborate with GATA1 at some genomic loci such as the locus control regions of beta-globin, while blocks GATA1 binding at other loci.

Together, these findings provide a potential mechanistic explanation for our lab's previous observation that CUX1 loss led to anemia and fatal myeloid malignancies in mice though disrupting the endogenous GATA1 binding. The novel interaction mechanism between two crucial erythropoiesis master regulators will provide key insights for therapeutic intervention.

CHAPTER 6. DISCUSSION

6.1 Overview

The loss of all chromosome 7 (-7) or its long arm (del(7q)) is one of the most recognized high risk cytogenetic abnormalities in pediatric and adult myeloid malignancies.[29] Despite the clinical implication of -7/del(7q), the underlying mechanism that promotes transformation has remained elusive. It is now recognized that cooperation of multiple haploinsufficient 7q genes, rather than complete inactivation of a 7q TSG, are likely responsible for disease progression. However, identifying the critical genes has been difficult given the large number of genes in the deleted regions. The first project of this thesis leveraged machine learning to providing a comprehensive tool for identifying del 7q TSGs. By systematically mining publicly available genome-wide perturbation screens, the project provided a framework to discover tumor suppressor properties at a large scale by learning the behaviors of known TSGs from the phenotype of cellular proliferation. This work resulted in a ranked list of all human chromosome 7 genes with high-to-low TSG likeness scores, which can serve as a starting reference for future experiments.

The del(7q) gene, CUX1, had been shown by our lab and others to function as a critical myeloid TSG. The question remains: how does haploinsufficient expression of this homeobox-containing TF lead to malignancies? Myeloid neoplasms are characterized by disruptions to the myeloid differentiation process. We find that CUX1 possesses PF activities to epigenetically regulate hematopoietic lineage commitment and homeostasis. Herein, we demonstrate that CUX1 directs the BAF chromatin remodeling complex to DNA to increase chromatin accessibility in hematopoietic cells. CUX1 preferentially regulates lineage-specific enhancers, and CUX1 target genes are predictive of cell fate *in vivo*. Since del(7q) patients and Cux1-deficient mice both develop severe anemia, we finally examine the erythroid

branch of hematopoiesis. We find that CUX1 promotes healthy differentiation through ensuring proper GATA1 binding, a critical regulator of erythropoiesis.

Overall, the thesis offered mechanistic insights into myeloid malignancies associated with -7/del(7q), providing a comprehensive tool for identifying TSGs, clarifying the role of CUX1 as a PF in HSPC fate determination, and discovering an unexpected interaction mechanism between two major erythropoiesis regulators GATA1 and CUX1. These findings offer key mechanistic insights that underlie phenotypes observed in myeloid cancer patient.

6.2 Machine learning models uncover TSG activities that are often elusive to conventional analysis.

By applying a supervised learning approach using the random forest algorithm, we utilized the abundant genome-wide perturbation studies in human hematopoietic cancer cell lines, and predicted TSG activities systemically on the human chromosome 7. The main value proposition of my approach is that ML algorithms are good in detecting small, subtle but concordant changes across large-scale datasets. Identifying TSGs manually through simple statistical analysis one dataset at a time is not only time consuming, but also risks losing valuable false negative hits which not necessarily possess strong TSG activity in one screen but have weak yet real TSG activities across multiple screens. As the data source for such analysis became ever more abundant and our biological understanding for tumor suppressor advance. This classifier should only be a starting point rather than a wrapped-up project hidden on the bookshelf. There are multiple ways to improve the classifier.

To better characterize hematopoietic TSG activities, more screen data and better subtyping based on cancer types should be incorporated in the model training process. ML prediction results are only good as the input data. Databases such as BioGRID ORCS actively curated CRISPR-screens as they are published.[331] A search in the database revealed 50+ new CRISPR screens published in using human hematopoietic cell lines alone since the ending of this project. This number is only going to grow larger over time. Incorporating these results and re-running the classifier iteratively would increase the accuracy of TSG identification. Furthermore, as gene activity could differ across different cancer types, accurately subgrouping training data based on cancer cell types will help to pinpoint the context-dependent TSG activities more accurately. One of the major limitations of this study at the time is that the genome-wide genetic perturbation training data is limited, so we included several different types of cancer cell lines (**Table 2**). Although the majority of the cell lines are AML, several CML, ALL and solid tumor types including breast, melanoma and pancreatic cell lines were also included. With the ever-growing curated data, future implementation of the model should also train on different cancer types separately.

Secondly, the model training process should incorporate more modality/types of data source that can accurately represents TSG activity. TSG activity does not only manifest in cellular proliferation. DNA damage repair, post-translational modification, epigenetic modification, and metabolism are examples of other biological processes that could be affected by TSG.[278] All of these biological processes could manifest partially but not completely in proliferation. Therefore, multimodal data sourced reflecting these different biological aspects should be incorporated in predicting TSG activities. Indeed, DNA sequencing data which identify mutational patterns have been used to predict TSG activities.[42,332] A study has also identified histone modifications as strong predictor for

TSGs.[333] In recent years, mapping protein structure to biological function has become an intensively researched area, partially fuelled by advancement in deep learning-based Alphafold algorithm,[334] which could accurately predict 3D protein structure from primary amino acid sequences. Such mapping is significant because majority of the current therapeutic candidates work by targeting 3D protein structures, and finding the exact protein structure responsible for TSG activity could help cancer therapy development. A recent study applied convoluted neural network on 1,191 TSG and 1,188 OG protein structures from the protein data bank (PDB). The classifier is trained on extracted features including biochemical properties such as surface amino acid charge and hydrophobicity etc.[335] Such studies could help researchers to rapidly identify druggable target on protein surface to develop drugs that activate TSGs and inhibit cancer growth. In addition to the data sources mentioned so far, it is conceivable that data such as proteomics on post-translational modification, metabolomics, and epigenomes data such as genome-wide methylation will be invaluable training resources. Several studies have utilized these data to predict TSG activities.[279–282] However, to our knowledge, there are no studies to date that systemically integrating these multimodal data types to predict TSG activity in human cancer. It is certainly a very promising area to pursue in the future.

There is also a growing body of evidence suggesting that a subset of genes, often encoding for TFs and kinases, do not conform to the simplistic binary oncogene vs TSG categorization. Rather, they could either promote or inhibit cancer growth depending on cellular context.[284] Examples include the NOTCH receptors that functions as an oncogene in T-lineage acute lymphoblastic leukemia while it performs tumor-suppressor function in squamous epithelial cells.[336,337] Other "double agent" genes include PTP1B,[338] TP53[339] and WT1[340]. Cancer types, mutations and isoforms are the most common context-dependent

factors that dictates whether these “double agent” genes act as a tumor suppressors or oncogenes. Conceptually, it is possible that sequential mutational events first abolish the TSG activity of a gene and a second mutational events turn it oncogenic. These observations further strengthen the need to optimize the ML classifier to be more context driven.

6.3 CUX1 has pioneering factor activity in early hematopoiesis.

CUX1, a ubiquitously expressed TF, has emerged as a central player in hematopoietic malignancies. Haploinsufficiency of CUX1 disrupts normal HSPC homeostasis and differentiation, resulting in clonal expansion, lineage biases, and multilineage dysplasia.[79,341] When combined with additional mutations, CUX1 deficiency promotes fulminant leukemic transformation.[69,342] The role of CUX1 as a pioneer factor in regulating hematopoietic stem and progenitor cell (HSPC) homeostasis and differentiation provides valuable insights into the epigenetic regulation of stem cell fate. The second part of the thesis established CUX1 as a hematopoietic pioneer factor. CUX1 recruits the chromatin remodeler BAF complex to open enhancer and collaborate with other hematopoietic TFs to regulate HSC fate. This sequence of events is apparent in a substantial portion of CUX1 DNA binding sites, exemplified by the “direct model” of CUX1-dependent SMARCA4 recruitment and increased DNA accessibility (**Figures 4.2B and 4.3F**). We also observed a similar number of CUX1 binding events that were not required for SMARCA4 recruitment in the “indirect model” (**Figures 4.2B and 4.3G**). In this latter category of sites, SMARCA4 is potentially recruited via alternate transcription factors such as SP1,[287,320] whose motif is enriched at the indirect sites. CUX1 binding may be independent of or might follow BAF recruitment to these sites. It is not obvious why chromatin accessibility also decreases at these indirect sites after CUX1

knockdown (**Fig. 4.3G).** Perhaps CUX1 also promotes an open chromatin state by recruitment of histone acetyltransferases at these sites.[292,293] Alternatively, the partial nucleosome destabilization mediated by CUX1 alone enables other factors to bind and stabilize the more open chromatin state.

The mechanism of recruiting the chromatin remodeler BAF complex by CUX1 is pivotal for CUX1 pioneering activity in stem cells. Going forward, it will be interesting to map other potential elements of the sequential CUX1 recruitment events beyond the BAF complex. What other TFs CUX1 recruit as downstream events after BAF-induced chromatin opening? Does CUX1 recruit other chromatin remodelers to promote DNA accessibility? Although our co-IP mass spec data did not suggest apparent additional chromatin remodeler complexes interacting with CUX1, extensive investigations into various chromatin remodelers across different cellular and tissue systems are warranted.

Integrative analysis using RNA-seq has revealed that the regulation of DNA accessibility by CUX1 correlates with the transcriptomic impacts on myeloid expansion and erythroid differentiation inhibition following CUX1 knockdown (**Figure 4.6C**). Identifying the specific subset of CUX1 binding sites that influence gene expression regulation would be of great interest. Maresca et al. 2023 implemented an acute protein depletion system on PF SOX2 and conducted nascent transcription analysis, demonstrating that SOX2-dependent open DNA sites—not merely SOX2 binding sites—are highly indicative of SOX2-regulated gene expression.[343] This study implicates PF binding and chromatin opening are both required for activating gene expression. It is conceivable that this principal applies to the CUX1-dependent ATAC sites that are bound by CUX1. However, whether and how CUX1 regulates gene expression indirectly for the CUX1-depedent ATAC sites without direct CUX1 binding remains unknown. It would be intriguing to compare the gene regulatory

mechanisms between the CUX1 bound and non-bound CUX1-dependent ATAC-seq sites. It is likely that for the CUX1-depedent ATAC-seq sites not directly bound by CUX1, additional collaborating TFs and histone modifying enzymes are vital for regulating gene expression activity. How these elements are influenced by CUX1 dosage requires further experimental study. Additionally, discerning the immediate versus delayed gene expression effects of CUX1 knockdown is essential; The CRISPR knockout system we used assesses gene expression changes several days in culture post-CUX1 loss, potentially reflecting both immediate and prolonged effects. Our laboratory is in the process of refining a degron system, which, when used alongside nascent RNA-seq, could precisely identify the immediate changes in RNA expression regulated by CUX1.

As in any adult tissue, hematopoietic differentiation requires stem and progenitor cells to undergo epigenetic reprogramming to commission and decommission the appropriate enhancers while reorienting genomic architecture to implement the pertinent mature cell transcriptional program. To date, the central actors in this process in the apex of the hematopoietic hierarchy have remained unclear. With respect to chromatin remodelers in normal hematopoiesis, our mechanistic knowledge of these factors, including the BAF complex, remains incomplete. Regarding the TFs that direct these complexes, a few pioneer factors have been identified, but these have largely been described in cell lines or to act in downstream progenitors.[322] To our knowledge, CUX1 is the first transcription factor reported to have demonstrated pioneer factor activity in the early stages of differentiation, in primary human HSPCs.

Although not measured here, a logical extension of our finding is that CUX1 regulates chromatin accessibility in other tissue types. Given the wide-ranging role of CUX1 in the homeostasis of diverse tissues, it seems improbable that CUX1 only regulates a stereotypical

set of target genes. CUX1 is conceivably a more general regulator of enhancer receptivity to activation via ensuing lineage-specific TFs. In this paradigm, CUX1 is critical for initiating epigenetic remodeling in tissue-specific stem cells, and lineage-determining TFs drive subsequent differentiation.

In myeloid malignancies, developmental syndromes, and other developmental contexts, CUX1 has haploinsufficient phenotypes.[65,288] Likewise, mutations in the BAF complex are commonly heterozygous in cancer and developmental disorders.[137] It remains to be determined how CUX1 haploinsufficiency impacts genome-wide BAF recruitment and DNA accessibility. We observed a widespread reduction of SMARCA4 occupancy after knocking out CUX1 in K562 (**Figure 4.2A**). The subsequent location and function of the displaced BAF complex post-CUX1 knockout present an interesting line of inquiry. One possibility is that loss of one copy of CUX1 untethers a portion of BAF to enable promiscuous BAF recruitment to *de novo* sites via other interacting partners.[344] We did not convincingly identify such a "gain-of-function" effect in K562 cells, where few de novo SMARCA4 binding sites are acquired after CUX1 knockdown (**Figure 4.2A**). More likely, CUX1 or BAF complex haploinsufficiency leads to either partial or complete loss of regulation at a subset of target sites. It is probable that the untethered BAF complex returns to the nucleoplasm. Investigating the destiny of this liberated BAF complex is of significant interest. Questions about whether they remain unbound, are broken down, or recycled by cellular machinery are vital. This information could greatly influence therapeutic strategies involving BAF inhibitors in cancer treatment, especially considering that patients with CUX1 haploinsufficiency might display reduced BAF complex binding, possibly diminishing their sensitivity to BAF inhibitors. The development of biomarkers to identify patients with CUX1 haploinsufficiency who might benefit from BAF-targeted therapies would be crucial,

potentially including assays to detect levels of CUX1 or BAF complex activity within tumors. More likely, CUX1 or BAF complex haploinsufficiency leads to either partial or complete loss of regulation at a subset of target sites. The tools to precisely address this important question and characterize dose-dependent binding sites for future studies are only recently emerging.

The canonical model of pioneer factor activity posits that after DNA accessibility is increased, "settler" TFs can subsequently bind DNA and execute gene expression. Indeed, we find the motifs and TF occupancy of several key regulators of hematopoietic differentiation uncovered at sites regulated by CUX1 and BAF. Obstensibly counterintuitively, several of these TFs independently harbor pioneer factor activity, including RUNX1, PU.1, and KLF1.[322] There are several potential explanations for this apparent redundancy. First, it is conceivable that more than one pioneer TF binds simultaneously to an enhancer to cooperatively establish the enhancer landscape during differentiation.[345] Second, and not mutually exclusive, a given TF does not necessarily have pioneer activity at all DNA targets, as we observed for CUX1 and was described for PU.1, as two examples.[157] In other words, CUX1 may be required for PU.1 binding at a subset of enhancers. It is conceivable that CUX1 and other PFs/non-PFs might pioneering for each other and form an intricate regulatory circuit in coordinating stage-specific gene expression, similar to the mutual pioneering action of FOXA1 and HNF4A.[120] Identifying such "co-pioneers" of CUX1 will be of great interest to understand the mechanism of CUX1 gene expression regulation. A third possibility is that these factors are binding sequentially, as opposed to simultaneously, during differentiation. In this case, CUX1 is required in HSPCs while a subsequent pioneer factor maintains accessibility in more mature progenitors. Thus, while the pioneer model provides a framework for conceptualizing epigenetic regulation, like many biological models, there is

likely more underlying complexity. In fact, the binary concept of pioneer vs. settler TFs has been drawn into question, and more TFs may be uncovered within a spectrum of pioneer-like activity.

On the translational side, our research presents several promising therapeutic implications. Considering that CUX1 often functions as a haploinsufficient tumor suppressor gene (TSG) in multiple cancer types and is commonly lost via -7/del(7q) in hematological malignancies,[341] restoring wild-type CUX1 expression stands out as a promising therapeutic strategy. Overexpressing or re-introducing lost TSG in cancer cells have been studied extensively as a potential avenue of cancer therapeutics.[346] The reintroduction of CUX1 could restore normal BAF complex DNA binding activity, potentially rectifying the aberrant gene expression profiles characteristic of CUX1 haploinsufficient cancers. To achieve this, gene therapy and genome editing present viable approaches for reinstating CUX1 expression. Techniques such as lentiviral vectors or electroporation could reintroduce CUX1 into *ex vivo* autologous HSC population, which could then be reinfused into patients. *In vivo* gene therapy could utilize AAV to deliver functional CUX1 cDNA. CRISPR-based gene editing might target bone marrow directly, facilitating the repair or replacement of the mutated CUX1 allele. The recent approval of the world's first CRISPR-based gene editing therapy Casgevy® for beta-thalassemia, which edits *BCL11A* enhancers in HSCs to modulate the fetal vs. adult hemoglobin production, underscores the potential of this approach. Similar CRISPR technologies, possibly combined with homology-directed repair or base editing, could correct the CUX1 mutation via whole DNA segment replacement or single nucleotide editing. Before implementing such therapies, however, critical aspects must be clarified. Our laboratory's research indicates that CUX1 influences cell fate in a dosage-dependent manner,[79] making precise dosage control crucial. Traditional gene therapies that knock in genes might lead to

overexpression and dosage-related toxicity. In contrast, epigenome editors offer the advantage of modulating gene expression levels with tuneable precision.[347] This technology uses a DNA-binding domain, such as deactivated Cas9 (dCas9), combined with effector proteins like KRAB, DNMT3, or VP64, to modulate gene expression via transcriptional regulation or DNA methylation/acetylation. Given CUX1's haploinsufficiency, epigenome editing is particularly well-suited because it can fine-tune the endogenous expression level of the gene—neither too little nor too much. Upregulating CUX1 expression could be achieved through acetylating CUX1 promoter or VP64-mediated transcription machinery recruitment. Epigenome editing has already been explored in animal models and early clinical studies for diseases caused by haploinsufficiency, such as Dravet syndrome.[348] It is noteworthy that CUX1 level decrease with age, as are HSC self-renewal capability and -7/del(7) myeloid neoplasms.[349,350] Envisioning a future where doctors might replenish CUX1 levels to promote longevity is not entirely far-fetched. However, before such advancements become reality, numerous technical and biological hurdles must be overcome. Challenges such as targeted delivery to specific tissues, mitigating off-target effects and immunotoxicity linked with genome and epigenome editing must be addressed. Nevertheless, these therapeutic prospects hold undeniable promise and signify a frontier in precision medicine.

Our research also shed light on other potential therapeutic concepts. Small molecule modulators could be designed to enhance the activity of the residual BAF complex left upon CUX1 loss, possibly by promoting the assembly and stability amplifying its chromatin remodeling function to offset the reduced CUX1 levels. Although just a small fraction, the gained BAF binding events upon knocking out CUX1 warrants further investigation (**Figure 4.2A**). For example, are some of these gained BAF binding due to recruitment by other oncogenic TFs? In this regard, inhibitors that prevent oncogenic transcription factors from

aberrantly interacting with the BAF complex in the absence of CUX1 would also be beneficial, redirecting the complex to its regular targets. Research at our lab is also trying to use synthetic lethality screens with CRISPR-Cas9 to identify genes that, when inhibited, selectively kill cancer cells with CUX1 haploinsufficiency. Another avenue might involve molecules that can recruit alternative chromatin remodeling complexes and acetyltransferases to sites typically targeted by CUX1-BAF, offering a compensatory mechanism for CUX1 loss. These therapeutic strategies would need to be carefully tailored to preserve the essential functions of the BAF complex in normal cells while selectively targeting the cancerous pathways, thereby optimizing the treatment of cancers characterized by CUX1 loss. Lastly, extensive research has established BAF complex role in promoting genome stability and facilitate DNA damage repair.[351] Based on our lab's finding that CUX1 also possess similar capabilities,[81] it will be interesting to investigate whether CUX1 and the BAF complex cooperate to facilitate DNA repair, or they act through independent pathways. This knowledge could provide novel strategies to develop DDR therapeutics.

6.4 CUX1 interacts with the erythropoiesis master regulator GATA1.

One of the key features of pioneer factors is their ability to recruit or cooperate with other TFs to regulate lineage-specific gene expression. Since CUX1 is indispensable for healthy erythropoiesis, an exciting question to investigate is how CUX1 collaborate with other erythroid TFs in co-ordinately regulating erythropoiesis. The third part of my thesis unveiled how CUX1 interacts with the erythropoiesis master regulator GATA1. Instead of recruiting GATA1, surprisingly we discovered that CUX1 functions as a gatekeeper to prevent abnormal GATA1 binding. This might explain why GATA1 is not observed as an interaction partner with CUX1 in our co-IP mass spec data.

The gatekeeper paradigm of interaction between CUX1 and GATA1 is novel. TFs frequently interact with each other to regulate lineage specific-gene expression. In this regard, there are multiple known ways TFs could interact with each other. First, pioneer TFs could recruit other TFs upon initial binding to DNA, as discussed in the second part of this thesis. TFs could also act synergistically, such as FOXP3 and ETS1 in regulating the regulatory T cell differentiation.[352] Alternatively, TFs could also antagonize each other. A classic example is STAT3 and STAT5 competitive binding in regulating the cytokine *IL17* expression during the regulatory T cell maturation process.[353] Another example is the classical PU.1 – GATA1 antagonism in hematopoiesis that regulates erythroid vs. myeloid lineage balance, where these two lineage-specific TFs antagonize each other by direct physical interaction, competitive DNA binding and transcriptional repression of each other.[207,354] A study on PU.1 also showed that TF can not only regulate gene expression by direct binding, but also by "stealing" partner TFs from their endogenous sites.[355] We reason that the "gatekeeping" interaction of CUX1 and GATA1 share some similarities with the TFs competing for common binding sites. However, unlike examples such as GATA1-PU.1 competition, which lead to two distinct cell fates, CUX1 and GATA1 both function to promote healthy erythropoiesis. Additional, unlike GATA1-PU.1 and the GATA1-GATA2 duals where one TF directly suppress the expression of another TF, we did not observe a meaningful increase of GATA1 RNA or protein expression upon CUX1 loss in K562. In this context, CUX1 is more likely serving as a "steer" to direct existing GATA1 protein binding and serving as a "shield" to prevent abnormal binding events rather than engaging in direct competition for binding sites. To our knowledge, there is no existing literature supporting the notion that two master TFs of the same lineage function in a similar manner, thus underscoring the novelty and significance of our discovery.

Human genome contains millions of copies of GATA motif, all of which could house GATA1 factor binding theoretically.[196] Therefore, it is conceivable that multiple regulatory mechanisms exist to ensure GATA1 bind to correct subset of loci. CUX1 is central in such mechanism because knocking out CUX1 "unleashed" GATA1 for promiscuous binding. Interestingly, by integrating with RNA-seq analysis, we discovered that these *de novo* GATA1 binding sites correspond to increased cellular proliferation and division and impaired erythroid differentiation processes (**Figure 5.2E**), which might explain the phenotypical observation that CUX1 loss leads to impaired erythropoiesis. A more detailed investigation into the individual genes among the *de novo* GATA1 sites could provide novel insights.

GATA1 influences gene expression by interacting with a myriad of co-activators like LMO2 and LDB1, and co-repressors such as ETO2 and the NuRD complex.[356,357] These elements play opposing roles in the organization of chromatin, with the equilibrium of their activities determining the extent of gene expression regulated by GATA1. The mechanism behind GATA1's selection of specific factors for recruitment to specific genomic loci is a subject of significant research interest, yet it is not fully comprehended. How CUX1 influences GATA1 to recruit other co-factor warrants further investigation. For example, how CUX1 and GATA1 collaboratively interact with the BAF complex? Part two of this thesis demonstrates that CUX1 promotes DNA accessibility by recruiting the BAF complex and promote enhancer accessibility. GATA1 has also been shown to recruit the BAF complex to DNA, where the BAF complex promotes GATA-1-dependent chromatin looping and transcriptional activation of α- and β-globin loci.[162,165] This process is essential to safeguard healthy erythropoiesis.[358] In addition, ample evidence suggests that the BAF complex is essential in promoting healthy erythropoiesis. Mouse embryos with mutated SMARCA4 display blocked erythroid differentiation, anemia and eventually die from the mutation.[161] Mechanistically, the chromatin remodelling, histone acetylation, DNA methylation and

transcription are disrupted at the beta globin locus.[161] SMARCA4 deletion in mouse model has also shown that the BAF complex is required for primitive erythropoiesis and vascular development.[358] Given the importance of the BAF complex in erythropoiesis and the large number of overlapping binding site of CUX1 and GATA1, going forward, it will be interesting to untangle how these two transcription factors cooperate to recruit the BAF complex, or do they compete for BAF binding at certain genome loci.

On the genome-wide scale, even though loss of CUX1 predominantly leads to chromatin closing, there is a minor section of the genome that becomes more accessible (**Figure 4.3A**). We discovered that these sites are selectively enriched for GATA motif and display a higher magnitude of GATA1 binding increase after CUX1 loss (**Figure 5.2F**). The BAF complex subunit ARID1B has been shown to regulates GATA1 gene expression and binding site accessibility.[359] This implies that perhaps in the sites where both DNA accessibility and GATA1 binding increased upon CUX1 loss, GATA1 might recruit or take in some untethered BAF complex to promote DNA accessibility. These sites also have increased active histone mark H3K27ac and enhancer mark H3K4me1 signals after knocking down CUX1 (**Figure 5.2C**). This is surprising because CUX1 overall promotes the active chromatin mark H3K27ac and H3K4me1 signal genome-wide (**Figure 4.3C**). Previous observation suggests that that CUX1 might recruit histone acetyltransferases (HAT) to deposit H3K27ac.[292,293] Perhaps CUX1 differentially regulates the accessibility and histone modification on different sections of the genome. Despite the genome-wide function, loss of CUX1 selectively opens these regions and create a permissive histone environment to allow GATA1 factor to come in and bind. Alternatively, GATA1 binding itself could be the reason why these sites display permissive chromatin characteristics, since GATA1 has pioneer capabilities and is known to promote histone acetylation by recruiting HAT.[180,360,361] A plausible possibility for this seemingly contradictory observation is maybe the activating

capability of GATA1 surpassed the loss of DNA accessibility and histone acetylation caused by CUX1 loss. Future studies should vigorously investigate these alternative mechanisms and determine the sequence of such events.

CUX1 gatekeeps GATA1 binding by both direct shielding and unknown indirect mechanisms (**Figure 5.2A**). Future studies should aim to discover what these indirect mechanisms are. For example, CUX1 could collaborate with GATA1-antaganizing TFs such as PU.1 or GATA2 which reside at these sites to block GATA1 binding. Or does CUX1 creates repressive histone environment at these sites to prevent GATA1 binding?

This study opens many other open questions worth further investigation. For example, how does CUX1 loss affect the "GATA switch", which is an essential step in early HSC to erythroid progenitor transition? How does CUX1 affect the assembly of the LDB1 complex which is formed by GATA1 and several other co-factors to promote erythropoiesis? Furthermore, despite the overwhelming gatekeeping function, there are also a few GATA1 binding sites that decreased in occupancy upon CUX1 loss (**Figure 5.1A**), future studies should examine the identity and functions of these sites. Are these important regulatory sites for pro-erythrocyte differentiation genes? One possibility is that besides gatekeeping, CUX1 also recruits GATA1 to promote selective erythroid lineage gene expression. Lastly, since CUX1 is essential in maintaining healthy erythropoiesis, does it also gatekeeps other erythropoietic TFs such as TAL1 and KLF1? Does similar gatekeeping mechanism happen in other tissue systems where CUX1 plays essential developmental roles? What other TFs than CUX1 can serve as a gatekeeper in development?

While the research offers significant insights into the molecular mechanisms governing CUX1-GATA1 interaction and its dysregulation in erythropoiesis, the conclusion we obtained uses the K562 cancer cell line which is multipotent human leukemic progenitors. Future studies should test the hypothesis across different stages in primary mice and human

erythroid progenitors. Contrary to the predominant pioneer-like chromatin accessibility-promoting role of CUX1 in K562 and human HSPC, CUX1 loss leads to vast chromatin opening in primary mice RII cells, which damages the normal chromatin condensation process. This seemingly contradictory observation could be explained by CUX1 dosage. It has been observed that pioneer factors could lose the pioneer activity and switch "mode" to binding open DNA regions with other TFs when expressed at low dosage.[121] We have observed that CUX1 expression is much lower in RII comparing to HSPCs through western blot. Therefore, it is conceivable that CUX1 might lose the pioneer activity in RII and switch to other functions. What these functions are warrant further investigation. For example, does CUX1 collaborate with proteins in the cytoskeleton organization and histone methylation families to ensure chromatin condensation? Furthermore, how these seemingly opposite roles of CUX1 in chromatin accessibility affect the GATA1 binding is worth further investigating. Does CUX1 interact with GATA1 differently across different erythropoiesis stages, if so, what are the phenotypical consequences at each stage?

Our research provided a mechanistic explanation on why losing CUX1 leads to impaired erythropoiesis. Therapeutics and diagnostics strategies could be developed accordingly. Both CUX1 haploinsufficiency/mutation and GATA1 mutation status could be used as biomarker for prognosis purposes. Perhaps patients with dual CUX1 & GATA1 mutations will have worse anemia and survival than patients with just one mutation? In AML/MDS patients with CUX1 haploinsufficiency and severe anemia, selectively inhibiting GATA1 DNA-binding domain function could be used with precaution on preserving some degree of normal GATA1 activity while reducing promiscuous GATA1 binding. Although not invented yet, we envision one day therapeutic methods could redirect TF binding from loci to loci. Given the haploinsufficient state of cancer patients and the genomic "zip code" of

where CUX1 gatekeeps GATA1 binding, such wonder drug might be able to redirect promiscuous GATA1 to where it is supposed to bind.

In conclusion, the research presented in this thesis provides a computational framework to uncover TSGs from rich genome-wide screen data, and comprehensive explored the molecular mechanisms governing hematopoiesis and its dysregulation in cancer. The research established CUX1 as a pioneer TF in hematopoiesis by recruiting chromatin remodeler BAF complex to open lineage-specific enhancers, drives cell fate gene expression programs, and safeguard against abnormal GATA1 to ensure proper erythroid differentiation. the findings provide a solid foundation for future research and contribute significantly to our understanding of the complex regulatory networks in hematopoietic stem cell differentiation.

6.5 Future directions

Enrich ML training with multi-modal data integration: Explore other data types to add to the genome-wide perturbation screens in order to capture TSG activities beyond proliferation. Proteomics data on partner proteins, metabolomics, structural data and biochemical assays testing for DNA repair capabilities can likely be incorporated to capture these important readouts and increase the accuracy of TSG prediction. The same ML workflow could be applied to study other recurrent aneuploidy events beyond -7/del(7q), such as del(5q), trisomy 8, and trisomy 12 etc.

Chromosome 7 TSG interactions: Which interaction partners on chromosome 7 does predicted TSGs interact with in HSPCs? Perturb-seq might be used to probe for chromosome 7 TSG regulatory circuitry in a massive parallel way and reveal such interactions in the heterogeneous HSPC population with single cell resolution.

Mapping sequential CUX1 recruitment events: Investigate which TFs are recruited by CUX1 following BAF-induced chromatin opening, and whether CUX1 recruits other chromatin remodelers to promote DNA accessibility. The research should examine different cellular and tissue systems.

CUX1 binding site characterization: Identify the specific subset of CUX1 binding sites influencing gene expression regulation. Analyze whether CUX1's regulatory effect on DNA accessibility correlates with the transcriptomic impacts on myeloid expansion and erythroid differentiation.

CUX1-dependent ATAC sites: Explore the gene regulatory mechanisms between CUX1-bound and non-bound CUX1-dependent ATAC-seq sites, considering the potential roles of collaborating TFs and histone-modifying enzymes in regulating gene expression activity.

Immediate vs. delayed effects of CUX1 knockdown: Differentiate the immediate versus delayed gene expression effects of CUX1 knockdown using a refined degron system alongside nascent RNA-seq.

Where the untethered BAF complex goes: Determine how CUX1 haploinsufficiency impacts genome-wide BAF recruitment and DNA accessibility, considering the location and function of the BAF complex post-CUX1 knockout.

Cooperative pioneering of CUX1: Identify potential "co-pioneers" of CUX1 and understand the complex regulatory network of stage-specific gene expression by studying the cooperative or sequential binding of pioneer TFs during hematopoietic differentiation.

Therapeutic implications of restoring CUX1: Investigate gene therapy and genome editing techniques to restore CUX1 expression, focusing on dosage control and the development of biomarkers to identify patients who might benefit from BAF-targeted therapies.

Gatekeeping function of CUX1 in GATA1 binding: Study the indirect mechanisms by which CUX1 gatekeeps GATA1 binding, and whether CUX1 collaborates with GATA1-antagonizing TFs or creates a repressive histone environment to prevent GATA1 binding.

CUX1 and erythropoiesis: Examine how CUX1 loss affects the GATA switch and the assembly of the LDB1 complex in erythropoiesis, and investigate the gatekeeping role of CUX1 on other erythropoietic TFs like TAL1 and KLF1.

Translational approaches for CUX1 and GATA1 mutations: Develop therapeutic and diagnostic strategies based on CUX1 haploinsufficiency/mutation and GATA1 mutation status as biomarkers for prognosis. Explore the potential of selectively inhibiting GATA1 DNA-binding domain function while preserving normal GATA1 activity to reduce promiscuous binding in patients with anemia.

Mechanistic studies across erythropoiesis stages: Investigate how CUX1 interacts with GATA1 across different erythropoiesis stages and what the phenotypical consequences are at each stage, using primary mice and human erythroid progenitors.

REFERENCES

1. Mitelman, F., Johansson, B., and Mertens, F. (2007). The impact of translocations and gene fusions on cancer causation. Nat. Rev. Cancer *7*, 233–245. 10.1038/nrc2091.

2. Nowell, P.C. (1962). The minute chromosome (Phl) in chronic granulocytic leukemia. Blut *8*, 65–66. 10.1007/BF01630378.

3. Boveri, T. (2008). Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. J. Cell Sci. *121 Suppl 1*, 1–84. 10.1242/jcs.025742.

4. Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhsng, C.-Z., Wala, J., Mermel, C.H., et al. (2013). Pan-cancer patterns of somatic copy number alteration. Nat. Genet. *45*, 1134–1140. 10.1038/ng.2760.

5. Holland, A.J., and Cleveland, D.W. (2009). Boveri revisited: Chromosomal instability, aneuploidy and tumorigenesis. Nat. Rev. Mol. Cell Biol. *10*, 478–487. 10.1038/nrm2718.

6. Compton, D.A. (2011). Mechanisms of aneuploidy. Curr. Opin. Cell Biol. *23*, 109–113. 10.1016/j.ceb.2010.08.007.

7. Gordon, D.J., Resio, B., and Pellman, D. (2012). Causes and consequences of aneuploidy in cancer. Nat. Rev. Genet. *13*, 189–203. 10.1038/nrg3123.

8. Barnard, D.R., Kalousek, D.K., Wiersma, S.R., Lange, B.J., Benjamin, D.R., Arthur, D.C., Buckley, J.D., Kobrinsky, N., Neudorf, S., Sanders, J., et al. (1996). Morphologic, immunologic, and cytogenetic classification of acute myeloid leukemia and myelodysplastic syndrome in childhood: a report from the Childrens Cancer Group. Leukemia *10*, 5–12.

9. Maurici, D., Perez-Atayde, A., Grier, H.E., Baldini, N., Serra, M., and Fletcher, J.A. (1998). Frequency and implications of chromosome 8 and 12 gains in Ewing sarcoma. Cancer Genet. Cytogenet. *100*, 106–110. 10.1016/s0165-4608(97)00028-9.

10. Weaver, B.A.A., and Cleveland, D.W. (2006). Does aneuploidy cause cancer? Curr. Opin. Cell Biol. *18*, 658–667. 10.1016/j.ceb.2006.10.002.

11. Taylor, A.M., Shih, J., Ha, G., Gao, G.F., Zhang, X., Berger, A.C., Schumacher, S.E., Wang, C., Hu, H., Liu, J., et al. (2018). Genomic and Functional Approaches to Understanding Cancer Aneuploidy. Cancer Cell *33*, 676-689.e3. 10.1016/j.ccell.2018.03.007.

12. Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., et al. (2010). The landscape of somatic copy-number alteration across human cancers. Nature *463*, 899–905. 10.1038/nature08822.

13. Williams, B.R., Prabhu, V.R., Hunter, K.E., Glazier, C.M., Whittaker, C.A., Housman, D.E., and Amon, A. (2008). Aneuploidy affects proliferation and spontaneous immortalization in mammalian cells. Science *322*, 703–709. 10.1126/science.1160058.

14. Torres, E.M., Sokolsky, T., Tucker, C.M., Chan, L.Y., Boselli, M., Dunham, M.J., and Amon, A. (2007). Effects of aneuploidy on cellular physiology and cell division in haploid yeast. Science *317*, 916–924. 10.1126/science.1142210.

15. Stingele, S., Stoehr, G., Peplowska, K., Cox, J., Mann, M., and Storchova, Z. (2012). Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. Mol. Syst. Biol. *8*, 608. 10.1038/msb.2012.40.

16. Ben-David, U., and Amon, A. (2020). Context is everything: aneuploidy in cancer. Nat. Rev. Genet. *21*, 44–62. 10.1038/s41576-019-0171-x.

17. Sheltzer, J.M., Ko, J.H., Replogle, J.M., Habibe Burgos, N.C., Chung, E.S., Meehl, C.M., Sayles, N.M., Passerini, V., Storchova, Z., and Amon, A. (2017). Single-chromosome Gains Commonly Function as Tumor Suppressors. Cancer Cell *31*, 240–255. 10.1016/j.ccell.2016.12.004.

18. Sansregret, L., and Swanton, C. (2017). The Role of Aneuploidy in Cancer Evolution. Cold Spring Harb. Perspect. Med. *7*, a028373. 10.1101/cshperspect.a028373.

19. Rutledge, S.D., Douglas, T.A., Nicholson, J.M., Vila-Casadesús, M., Kantzler, C.L., Wangsa, D., Barroso-Vilares, M., Kale, S.D., Logarinho, E., and Cimini, D. (2016). Selective advantage of trisomic human cells cultured in non-standard conditions. Sci. Rep. *6*, 22828. 10.1038/srep22828.

20. Baker, D.J., Jin, F., Jeganathan, K.B., and van Deursen, J.M. (2009). Whole chromosome instability caused by Bub1 insufficiency drives tumorigenesis through tumor suppressor gene loss of heterozygosity. Cancer Cell *16*, 475–486. 10.1016/j.ccr.2009.10.023.

21. Ricke, R.M., Jeganathan, K.B., and van Deursen, J.M. (2011). Bub1 overexpression induces aneuploidy and tumor formation through Aurora B kinase hyperactivation. J. Cell Biol. *193*, 1049–1064. 10.1083/jcb.201012035.

22. Bakhoum, S.F., Ngo, B., Laughney, A.M., Cavallo, J.-A., Murphy, C.J., Ly, P., Shah, P., Sriram, R.K., Watkins, T.B.K., Taunk, N.K., et al. (2018). Chromosomal instability drives metastasis through a cytosolic DNA response. Nature *553*, 467–472. 10.1038/nature25432.

23. Replogle, J.M., Zhou, W., Amaro, A.E., McFarland, J.M., Villalobos-Ortiz, M., Ryan, J., Letai, A., Yilmaz, O., Sheltzer, J., Lippard, S.J., et al. (2020). Aneuploidy increases resistance to chemotherapeutics by antagonizing cell division. Proc. Natl. Acad. Sci. U. S. A. *117*, 30566–30576. 10.1073/pnas.2009506117.

24. Ippolito, M.R., Martis, V., Martin, S., Tijhuis, A.E., Hong, C., Wardenaar, R., Dumont, M., Zerbib, J., Spierings, D.C.J., Fachinetti, D., et al. (2021). Gene copy-number changes and chromosomal instability induced by aneuploidy confer resistance to chemotherapy. Dev. Cell *56*, 2440-2454.e6. 10.1016/j.devcel.2021.07.006.

25. Lukow, D.A., Sausville, E.L., Suri, P., Chunduri, N.K., Wieland, A., Leu, J., Smith, J.C., Girish, V., Kumar, A.A., Kendall, J., et al. (2021). Chromosomal instability accelerates the evolution of resistance to anti-cancer therapies. Dev. Cell *56*, 2427-2439.e4. 10.1016/j.devcel.2021.07.009.

26. Hieronymus, H., Murali, R., Tin, A., Yadav, K., Abida, W., Moller, H., Berney, D., Scher, H., Carver, B., Scardino, P., et al. (2018). Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death. eLife *7*, e37294. 10.7554/eLife.37294.

27. Kawankar, N., and Vundinti, B.R. (2011). Cytogenetic abnormalities in myelodysplastic syndrome: an overview. Hematol. Amst. Neth. *16*, 131–138. 10.1179/102453311X12940641877966.

28. Rowley, J.D. (1973). Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature *243*, 290–293. 10.1038/243290a0.

29. Arber, D.A., Orazi, A., Hasserjian, R., Thiele, J., Borowitz, M.J., Le Beau, M.M., Bloomfield, C.D., Cazzola, M., and Vardiman, J.W. (2016). The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood *127*, 2391–2405. 10.1182/blood-2016-03-643544.

30. Rowley, J.D. (1973). Letter: Deletions of chromosome 7 in haematological disorders. Lancet Lond. Engl. *2*, 1385–1386. 10.1016/s0140-6736(73)93347-3.

31. Takahashi, K., Wang, F., Kantarjian, H., Song, X., Patel, K., Neelapu, S., Gumbs, C., Little, L., Tippen, S., Thornton, R., et al. (2017). Copy number alterations detected as clonal hematopoiesis of indeterminate potential. Blood Adv. *1*, 1031–1036. 10.1182/bloodadvances.2017007922.

32. Dimitriou, M., Woll, P.S., Mortera-Blanco, T., Karimi, M., Wedge, D.C., Doolittle, H., Douagi, I., Papaemmanuil, E., Jacobsen, S.E.W., and Hellström-Lindberg, E. (2016). Perturbed hematopoietic stem and progenitor cell hierarchy in myelodysplastic syndromes patients with monosomy 7 as the sole cytogenetic abnormality. Oncotarget *7*, 72685–72698. 10.18632/oncotarget.12234.

33. Luna-Fineman, S., Shannon, K.M., and Lange, B.J. (1995). Childhood monosomy 7: epidemiology, biology, and mechanistic implications. Blood *85*, 1985–1999.

34. Pezeshki, A., Podder, S., Kamel, R., and Corey, S.J. (2017). Monosomy 7/del (7q) in inherited bone marrow failure syndromes: A systematic review. Pediatr. Blood Cancer *64*. 10.1002/pbc.26714.

35. Dumitriu, B., Feng, X., Townsley, D.M., Ueda, Y., Yoshizato, T., Calado, R.T., Yang, Y., Wakabayashi, Y., Kajigaya, S., Ogawa, S., et al. (2015). Telomere attrition and candidate gene mutations preceding monosomy 7 in aplastic anemia. Blood *125*, 706–709. 10.1182/blood-2014-10-607572.

36. McNerney, M.E., Brown, C.D., Peterson, A.L., Banerjee, M., Larson, R.A., Anastasi, J., Le Beau, M.M., and White, K.P. (2014). The spectrum of somatic mutations in high-risk

acute myeloid leukaemia with -7/del(7q). Br. J. Haematol. *166*, 550–556. 10.1111/bjh.12964.

37. Smith, S.M., Le Beau, M.M., Huo, D., Karrison, T., Sobecks, R.M., Anastasi, J., Vardiman, J.W., Rowley, J.D., and Larson, R.A. (2003). Clinical-cytogenetic associations in 306 patients with therapy-related myelodysplasia and myeloid leukemia: the University of Chicago series. Blood *102*, 43–52. 10.1182/blood-2002-11-3343.

38. Jerez, A., Sugimoto, Y., Makishima, H., Verma, A., Jankowska, A.M., Przychodzen, B., Visconte, V., Tiu, R.V., O'Keefe, C.L., Mohamedali, A.M., et al. (2012). Loss of heterozygosity in 7q myeloid disorders: clinical associations and genomic pathogenesis. Blood *119*, 6109–6117. 10.1182/blood-2011-12-397620.

39. Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V.I., Paschka, P., Roberts, N.D., Potter, N.E., Heuser, M., Thol, F., Bolli, N., et al. (2016). Genomic Classification and Prognosis in Acute Myeloid Leukemia. N. Engl. J. Med. *374*, 2209–2221. 10.1056/NEJMoa1516192.

40. Schwartz, J.R., Ma, J., Lamprecht, T., Walsh, M., Wang, S., Bryant, V., Song, G., Wu, G., Easton, J., Kesserwan, C., et al. (2017). The genomic landscape of pediatric myelodysplastic syndromes. Nat. Commun. *8*, 1557. 10.1038/s41467-017-01590-5.

41. Knudson, A.G. (1971). Mutation and cancer: statistical study of retinoblastoma. Proc. Natl. Acad. Sci. U. S. A. *68*, 820–823. 10.1073/pnas.68.4.820.

42. Davoli, T., Xu, A.W., Mengwasser, K.E., Sack, L.M., Yoon, J.C., Park, P.J., and Elledge, S.J. (2013). Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell *155*, 948–962. 10.1016/j.cell.2013.10.011.

43. White, J.K., Gerdin, A.-K., Karp, N.A., Ryder, E., Buljan, M., Bussell, J.N., Salisbury, J., Clare, S., Ingham, N.J., Podrini, C., et al. (2013). Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. Cell *154*, 452–464. 10.1016/j.cell.2013.06.022.

44. Ernst, T., Chase, A.J., Score, J., Hidalgo-Curtis, C.E., Bryant, C., Jones, A.V., Waghorn, K., Zoi, K., Ross, F.M., Reiter, A., et al. (2010). Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. Nat. Genet. *42*, 722–726. 10.1038/ng.621.

45. Schmickel, R.D. (1986). Contiguous gene syndromes: a component of recognizable syndromes. J. Pediatr. *109*, 231–241. 10.1016/s0022-3476(86)80377-8.

46. Wong, J.C., Weinfurtner, K.M., Alzamora, M.D.P., Kogan, S.C., Burgess, M.R., Zhang, Y., Nakitandwe, J., Ma, J., Cheng, J., Chen, S.-C., et al. (2015). Functional evidence implicating chromosome 7q22 haploinsufficiency in myelodysplastic syndrome pathogenesis. eLife *4*, e07839. 10.7554/eLife.07839.

47. Stoddart, A., Wang, J., Fernald, A.A., Karrison, T., Anastasi, J., and Le Beau, M.M. (2014). Cell intrinsic and extrinsic factors synergize in mice with haploinsufficiency for

Tp53, and two human del(5q) genes, Egr1 and Apc. Blood *123*, 228–238. 10.1182/blood-2013-05-506568.

48. Xue, W., Kitzing, T., Roessler, S., Zuber, J., Krasnitz, A., Schultz, N., Revill, K., Weissmueller, S., Rappaport, A.R., Simon, J., et al. (2012). A cluster of cooperating tumor-suppressor gene candidates in chromosomal deletions. Proc. Natl. Acad. Sci. U. S. A. *109*, 8212–8217. 10.1073/pnas.1206062109.

49. McNerney, M.E., Brown, C.D., Wang, X., Bartom, E.T., Karmakar, S., Bandlamudi, C., Yu, S., Ko, J., Sandall, B.P., Stricker, T., et al. (2013). CUX1 is a haploinsufficient tumor suppressor gene on chromosome 7 frequently inactivated in acute myeloid leukemia. Blood *121*, 975–983. 10.1182/blood-2012-04-426965.

50. Döhner, K., Brown, J., Hehmann, U., Hetzel, C., Stewart, J., Lowther, G., Scholl, C., Fröhling, S., Cuneo, A., Tsui, L.C., et al. (1998). Molecular cytogenetic characterization of a critical region in bands 7q35-q36 commonly deleted in malignant myeloid disorders. Blood *92*, 4031–4035.

51. Chen, C., Liu, Y., Rappaport, A.R., Kitzing, T., Schultz, N., Zhao, Z., Shroff, A.S., Dickins, R.A., Vakoc, C.R., Bradner, J.E., et al. (2014). MLL3 is a haploinsufficient 7q tumor suppressor in acute myeloid leukemia. Cancer Cell *25*, 652–665. 10.1016/j.ccr.2014.03.016.

52. Hosono, N., Makishima, H., Jerez, A., Yoshida, K., Przychodzen, B., McMahon, S., Shiraishi, Y., Chiba, K., Tanaka, H., Miyano, S., et al. (2014). Recurrent genetic defects on chromosome 7q in myeloid neoplasms. Leukemia *28*, 1348–1351. 10.1038/leu.2014.25.

53. Kipreos, E.T., Lander, L.E., Wing, J.P., He, W.W., and Hedgecock, E.M. (1996). cul-1 is required for cell cycle exit in C. elegans and identifies a novel gene family. Cell *85*, 829–839. 10.1016/s0092-8674(00)81267-2.

54. Nagamachi, A., Matsui, H., Asou, H., Ozaki, Y., Aki, D., Kanai, A., Takubo, K., Suda, T., Nakamura, T., Wolff, L., et al. (2013). Haploinsufficiency of SAMD9L, an endosome fusion facilitator, causes myeloid malignancies in mice mimicking human diseases with monosomy 7. Cancer Cell *24*, 305–317. 10.1016/j.ccr.2013.08.011.

55. Whalen, S., Schreiber, J., Noble, W.S., and Pollard, K.S. (2022). Navigating the pitfalls of applying machine learning in genomics. Nat. Rev. Genet. *23*, 169–181. 10.1038/s41576-021-00434-9.

56. Singh, A., Thakur, N., and Sharma, A. (2016). A review of supervised machine learning algorithms. In 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1310–1315.

57. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions | Journal of Big Data | Full Text https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00444-8.

58. Chen, X., and Ishwaran, H. (2012). Random forests for genomic data analysis. Genomics *99*, 323–329. 10.1016/j.ygeno.2012.04.003.

59. Kourou, K., Exarchos, K.P., Papaloukas, C., Sakaloglou, P., Exarchos, T., and Fotiadis, D.I. (2021). Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis. Comput. Struct. Biotechnol. J. *19*, 5546–5555. 10.1016/j.csbj.2021.10.006.

60. Koh, D.-M., Papanikolaou, N., Bick, U., Illing, R., Kahn, C.E., Kalpathi-Cramer, J., Matos, C., Martí-Bonmatí, L., Miles, A., Mun, S.K., et al. (2022). Artificial intelligence and machine learning in cancer imaging. Commun. Med. *2*, 133. 10.1038/s43856-022-00199-0.

61. Rong Zeng, W., Soucie, E., Sung Moon, N., Martin-Soudant, N., Bérubé, G., Leduy, L., and Nepveu, A. (2000). Exon/intron structure and alternative transcripts of the CUTL1 gene. Gene *241*, 75–85. 10.1016/s0378-1119(99)00465-5.

62. Goulet, B., Baruch, A., Moon, N.-S., Poirier, M., Sansregret, L.L., Erickson, A., Bogyo, M., and Nepveu, A. (2004). A cathepsin L isoform that is devoid of a signal peptide localizes to the nucleus in S phase and processes the CDP/Cux transcription factor. Mol. Cell *14*, 207–219. 10.1016/s1097-2765(04)00209-6.

63. Goulet, B., Watson, P., Poirier, M., Leduy, L., Bérubé, G., Meterissian, S., Jolicoeur, P., and Nepveu, A. (2002). Characterization of a tissue-specific CDP/Cux isoform, p75, activated in breast tumor cells. Cancer Res. *62*, 6625–6633.

64. Krishnan, M., Senagolage, M.D., Baeten, J.T., Wolfgeher, D.J., Khan, S., Kron, S.J., and McNerney, M.E. (2022). Genomic studies controvert the existence of the CUX1 p75 isoform. Sci. Rep. *12*, 151. 10.1038/s41598-021-03930-4.

65. Ramdzan, Z.M., and Nepveu, A. (2014). CUX1, a haploinsufficient tumour suppressor gene overexpressed in advanced cancers. Nat. Rev. Cancer *14*, 673–682. 10.1038/nrc3805.

66. Truscott, M., Raynal, L., Wang, Y., Bérubé, G., Leduy, L., and Nepveu, A. (2004). The N-terminal region of the CCAAT displacement protein (CDP)/Cux transcription factor functions as an autoinhibitory domain that modulates DNA binding. J. Biol. Chem. *279*, 49787–49794. 10.1074/jbc.M409484200.

67. Li, S., Moy, L., Pittman, N., Shue, G., Aufiero, B., Neufeld, E.J., LeLeiko, N.S., and Walsh, M.J. (1999). Transcriptional repression of the cystic fibrosis transmembrane conductance regulator gene, mediated by CCAAT displacement protein/cut homolog, is associated with histone deacetylation. J. Biol. Chem. *274*, 7803–7815. 10.1074/jbc.274.12.7803.

68. Aly, M., Ramdzan, Z.M., Nagata, Y., Balasubramanian, S.K., Hosono, N., Makishima, H., Visconte, V., Kuzmanovic, T., Adema, V., Nazha, A., et al. (2019). Distinct clinical and biological implications of CUX1 in myeloid neoplasms. Blood Adv. *3*, 2164–2178. 10.1182/bloodadvances.2018028423.

69. Wong, C.C., Martincorena, I., Rust, A.G., Rashid, M., Alifrangis, C., Alexandrov, L.B., Tiffen, J.C., Kober, C., Chronic Myeloid Disorders Working Group of the International

Cancer Genome Consortium, Green, A.R., et al. (2014). Inactivating CUX1 mutations promote tumorigenesis. Nat. Genet. *46*, 33–38. 10.1038/ng.2846.

70. Yoshizato, T., Dumitriu, B., Hosokawa, K., Makishima, H., Yoshida, K., Townsley, D., Sato-Otsubo, A., Sato, Y., Liu, D., Suzuki, H., et al. (2015). Somatic Mutations and Clonal Hematopoiesis in Aplastic Anemia. N. Engl. J. Med. *373*, 35–47. 10.1056/NEJMoa1414799.

71. Robertson, N.A., Latorre-Crespo, E., Terradas-Terradas, M., Lemos-Portela, J., Purcell, A.C., Livesey, B.J., Hillary, R.F., Murphy, L., Fawkes, A., MacGillivray, L., et al. (2022). Longitudinal dynamics of clonal hematopoiesis identifies gene-specific fitness effects. Nat. Med. *28*, 1439–1446. 10.1038/s41591-022-01883-3.

72. Sinclair, A.M., Lee, J.A., Goldstein, A., Xing, D., Liu, S., Ju, R., Tucker, P.W., Neufeld, E.J., and Scheuermann, R.H. (2001). Lymphoid apoptosis and myeloid hyperplasia in CCAAT displacement protein mutant mice. Blood *98*, 3658–3667. 10.1182/blood.v98.13.3658.

73. Tufarelli, C., Fujiwara, Y., Zappulla, D.C., and Neufeld, E.J. (1998). Hair defects and pup loss in mice with targeted deletion of the first cut repeat domain of the Cux/CDP homeoprotein gene. Dev. Biol. *200*, 69–81. 10.1006/dbio.1998.8950.

74. Luong, M.X., van der Meijden, C.M., Xing, D., Hesselton, R., Monuki, E.S., Jones, S.N., Lian, J.B., Stein, J.L., Stein, G.S., Neufeld, E.J., et al. (2002). Genetic ablation of the CDP/Cux protein C terminus results in hair cycle defects and reduced male fertility. Mol. Cell. Biol. *22*, 1424–1437. 10.1128/MCB.22.5.1424-1437.2002.

75. Ellis, T., Gambardella, L., Horcher, M., Tschanz, S., Capol, J., Bertram, P., Jochum, W., Barrandon, Y., and Busslinger, M. (2001). The transcriptional repressor CDP (Cutl1) is essential for epithelial cell differentiation of the lung and the hair follicle. Genes Dev. *15*, 2307–2319. 10.1101/gad.200101.

76. Lievens, P.M., Tufarelli, C., Donady, J.J., Stagg, A., and Neufeld, E.J. (1997). CASP, a novel, highly conserved alternative-splicing product of the CDP/cut/cux gene, lacks cut-repeat and homeo DNA-binding domains, and interacts with full-length CDP in vitro. Gene *197*, 73–81. 10.1016/s0378-1119(97)00243-6.

77. Gillingham, A.K., Pfeifer, A.C., and Munro, S. (2002). CASP, the alternatively spliced product of the gene encoding the CCAAT-displacement protein transcription factor, is a Golgi membrane protein related to giantin. Mol. Biol. Cell *13*, 3761–3774. 10.1091/mbc.e02-06-0349.

78. Lowe, M. (2019). The Physiological Functions of the Golgin Vesicle Tethering Proteins. Front. Cell Dev. Biol. *7*, 94. 10.3389/fcell.2019.00094.

79. An, N., Khan, S., Imgruet, M.K., Gurbuxani, S.K., Konecki, S.N., Burgess, M.R., and McNerney, M.E. (2018). Gene dosage effect of CUX1 in a murine model disrupts HSC homeostasis and controls the severity and mortality of MDS. Blood *131*, 2682–2697. 10.1182/blood-2017-10-810028.

80. Imgruet, M.K., Lutze, J., An, N., Hu, B., Khan, S., Kurkewich, J., Martinez, T.C., Wolfgeher, D., Gurbuxani, S.K., Kron, S.J., et al. (2021). Loss of a 7q gene, CUX1, disrupts epigenetically driven DNA repair and drives therapy-related myeloid neoplasms. Blood *138*, 790–805. 10.1182/blood.2020009195.

81. Imgruet, M.K., Lutze, J., An, N., Hu, B., Khan, S., Kurkewich, J., Martinez, T.C., Wolfgeher, D., Gurbuxani, S.K., Kron, S.J., et al. (2021). Loss of a 7q gene, CUX1, disrupts epigenetically driven DNA repair and drives therapy-related myeloid neoplasms. Blood *138*, 790–805. 10.1182/blood.2020009195.

82. Kedinger, V., Sansregret, L., Harada, R., Vadnais, C., Cadieux, C., Fathers, K., Park, M., and Nepveu, A. (2009). p110 CUX1 homeodomain protein stimulates cell migration and invasion in part through a regulatory cascade culminating in the repression of E-cadherin and occludin. J. Biol. Chem. *284*, 27701–27711. 10.1074/jbc.M109.031849.

83. Cadieux, C., Kedinger, V., Yao, L., Vadnais, C., Drossos, M., Paquet, M., and Nepveu, A. (2009). Mouse mammary tumor virus p75 and p110 CUX1 transgenic mice develop mammary tumors of various histologic types. Cancer Res. *69*, 7188–7197. 10.1158/0008-5472.CAN-08-4899.

84. Cadieux, C., Fournier, S., Peterson, A.C., Bédard, C., Bedell, B.J., and Nepveu, A. (2006). Transgenic mice expressing the p75 CCAAT-displacement protein/Cut homeobox isoform develop a myeloproliferative disease-like myeloid leukemia. Cancer Res. *66*, 9492–9501. 10.1158/0008-5472.CAN-05-4230.

85. Zakrzewski, W., Dobrzyński, M., Szymonowicz, M., and Rybak, Z. (2019). Stem cells: past, present, and future. Stem Cell Res. Ther. *10*, 68. 10.1186/s13287-019-1165-5.

86. Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell *172*, 650–665. 10.1016/j.cell.2018.01.029.

87. Lee, T.I., and Young, R.A. (2013). Transcriptional Regulation and its Misregulation in Disease. Cell *152*, 1237–1251. 10.1016/j.cell.2013.02.014.

88. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321–1325. 10.1126/science.1098119.

89. Wunderlich, Z., and Mirny, L.A. (2009). Different gene regulation strategies revealed by analysis of binding motifs. Trends Genet. TIG *25*, 434–440. 10.1016/j.tig.2009.08.003.

90. Reiter, F., Wienerroither, S., and Stark, A. (2017). Combinatorial function of transcription factors and cofactors. Curr. Opin. Genet. Dev. *43*, 73–81. 10.1016/j.gde.2016.12.007.

91. Stadhouders, R., Filion, G.J., and Graf, T. (2019). Transcription factors and 3D genome conformation in cell-fate decisions. Nature *569*, 345–354. 10.1038/s41586-019-1182-7.

92. Ong, C.-T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nat. Rev. Genet. *15*, 234–246. 10.1038/nrg3663.

93. Briggs, M.R., Kadonaga, J.T., Bell, S.P., and Tjian, R. (1986). Purification and biochemical characterization of the promoter-specific transcription factor, Sp1. Science *234*, 47–52. 10.1126/science.3529394.

94. Kaczynski, J., Cook, T., and Urrutia, R. (2003). Sp1- and Krüppel-like transcription factors. Genome Biol. *4*, 206. 10.1186/gb-2003-4-2-206.

95. Porcher, C., Swat, W., Rockwell, K., Fujiwara, Y., Alt, F.W., and Orkin, S.H. (1996). The T cell leukemia oncoprotein SCL/tal-1 is essential for development of all hematopoietic lineages. Cell *86*, 47–57. 10.1016/s0092-8674(00)80076-8.

96. Gregory, T., Yu, C., Ma, A., Orkin, S.H., Blobel, G.A., and Weiss, M.J. (1999). GATA-1 and erythropoietin cooperate to promote erythroid cell survival by regulating bcl-xL expression. Blood *94*, 87–96.

97. Lyons, S.E., Lawson, N.D., Lei, L., Bennett, P.E., Weinstein, B.M., and Liu, P.P. (2002). A nonsense mutation in zebrafish gata1 causes the bloodless phenotype in vlad tepes. Proc. Natl. Acad. Sci. U. S. A. *99*, 5454–5459. 10.1073/pnas.082695299.

98. Pevny, L., Simon, M.C., Robertson, E., Klein, W.H., Tsai, S.F., D'Agati, V., Orkin, S.H., and Costantini, F. (1991). Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. Nature *349*, 257–260. 10.1038/349257a0.

99. Fujiwara, Y., Browne, C.P., Cunniff, K., Goff, S.C., and Orkin, S.H. (1996). Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor GATA-1. Proc. Natl. Acad. Sci. U. S. A. *93*, 12355–12358. 10.1073/pnas.93.22.12355.

100. Fisher, R.C., and Scott, E.W. (1998). Role of PU.1 in hematopoiesis. Stem Cells Dayt. Ohio *16*, 25–37. 10.1002/stem.160025.

101. Suh, H.C., Gooya, J., Renn, K., Friedman, A.D., Johnson, P.F., and Keller, J.R. (2006). C/EBPα determines hematopoietic cell fate in multipotential progenitor cells by inhibiting erythroid differentiation and inducing myeloid differentiation. Blood *107*, 4308–4316. 10.1182/blood-2005-06-2216.

102. van der Meer, L.T., Jansen, J.H., and van der Reijden, B.A. (2010). Gfi1 and Gfi1b: key regulators of hematopoiesis. Leukemia *24*, 1834–1843. 10.1038/leu.2010.195.

103. Duncliffe, K.N., Bert, A.G., Vadas, M.A., and Cockerill, P.N. (1997). A T cell-specific enhancer in the interleukin-3 locus is activated cooperatively by Oct and NFAT elements within a DNase I-hypersensitive site. Immunity *6*, 175–185. 10.1016/s1074-7613(00)80424-0.

104. Strubin, M., Newell, J.W., and Matthias, P. (1995). OBF-1, a novel B cell-specific coactivator that stimulates immunoglobulin promoter activity through association with octamer-binding proteins. Cell *80*, 497–506. 10.1016/0092-8674(95)90500-6.

105. Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet. *15*, 272–286. 10.1038/nrg3682.

106. Visel, A., Rubin, E.M., and Pennacchio, L.A. (2009). Genomic views of distant-acting enhancers. Nature *461*, 199–205. 10.1038/nature08451.

107. Stadhouders, R., van den Heuvel, A., Kolovos, P., Jorna, R., Leslie, K., Grosveld, F., and Soler, E. (2012). Transcription regulation by distal enhancers: who's in the loop? Transcription *3*, 181–186. 10.4161/trns.20720.

108. Mohrs, M., Blankespoor, C.M., Wang, Z.E., Loots, G.G., Afzal, V., Hadeiba, H., Shinkai, K., Rubin, E.M., and Locksley, R.M. (2001). Deletion of a coordinate regulator of type 2 cytokine expression in mice. Nat. Immunol. *2*, 842–847. 10.1038/ni0901-842.

109. Heinz, S., Romanoski, C.E., Benner, C., and Glass, C.K. (2015). The selection and function of cell type-specific enhancers. Nat. Rev. Mol. Cell Biol. *16*, 144–154. 10.1038/nrm3949.

110. Cirillo, L.A., Lin, F.R., Cuesta, I., Friedman, D., Jarnik, M., and Zaret, K.S. (2002). Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. Mol. Cell *9*, 279–289. 10.1016/s1097-2765(02)00459-8.

111. Cirillo, L.A., and Zaret, K.S. (1999). An early developmental transcription factor complex that is more stable on nucleosome core particles than on free DNA. Mol. Cell *4*, 961–969. 10.1016/s1097-2765(00)80225-7.

112. Balsalobre, A., and Drouin, J. (2022). Pioneer factors as master regulators of the epigenome and cell fate. Nat. Rev. Mol. Cell Biol. *23*, 449–464. 10.1038/s41580-022-00464-z.

113. Adams, E.J., Karthaus, W.R., Hoover, E., Liu, D., Gruet, A., Zhang, Z., Cho, H., DiLoreto, R., Chhangawala, S., Liu, Y., et al. (2019). FOXA1 mutations alter pioneering activity, differentiation and prostate cancer phenotypes. Nature *571*, 408–412. 10.1038/s41586-019-1318-9.

114. Jozwik, K.M., and Carroll, J.S. (2012). Pioneer factors in hormone-dependent cancers. Nat. Rev. Cancer *12*, 381–385. 10.1038/nrc3263.

115. Sun, Y., Zhou, B., Mao, F., Xu, J., Miao, H., Zou, Z., Phuc Khoa, L.T., Jang, Y., Cai, S., Witkin, M., et al. (2018). HOXA9 Reprograms the Enhancer Landscape to Promote Leukemogenesis. Cancer Cell *34*, 643-658.e5. 10.1016/j.ccell.2018.08.018.

116. Sunkel, B.D., Wang, M., LaHaye, S., Kelly, B.J., Fitch, J.R., Barr, F.G., White, P., and Stanton, B.Z. (2021). Evidence of pioneer factor activity of an oncogenic fusion transcription factor. iScience *24*, 102867. 10.1016/j.isci.2021.102867.

117. McPherson, C.E., Shim, E.Y., Friedman, D.S., and Zaret, K.S. (1993). An active tissue-specific enhancer and bound transcription factors existing in a precisely positioned nucleosomal array. Cell *75*, 387–398. 10.1016/0092-8674(93)80079-t.

118. Shim, E.Y., Woodcock, C., and Zaret, K.S. (1998). Nucleosome positioning by the winged helix transcription factor HNF3. Genes Dev. *12*, 5–10. 10.1101/gad.12.1.5.

119. Cirillo, L.A., McPherson, C.E., Bossard, P., Stevens, K., Cherian, S., Shim, E.Y., Clark, K.L., Burley, S.K., and Zaret, K.S. (1998). Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome. EMBO J. *17*, 244–254. 10.1093/emboj/17.1.244.

120. Hansen, J.L., Loell, K.J., and Cohen, B.A. (2022). A test of the pioneer factor hypothesis using ectopic liver gene activation. eLife *11*, e73358. 10.7554/eLife.73358.

121. Blassberg, R., Patel, H., Watson, T., Gouti, M., Metzis, V., Delás, M.J., and Briscoe, J. (2022). Sox2 levels regulate the chromatin occupancy of WNT mediators in epiblast progenitors responsible for vertebrate body formation. Nat. Cell Biol. *24*, 633–644. 10.1038/s41556-022-00910-2.

122. Orkin, S.H., and Zon, L.I. (2008). Hematopoiesis: An Evolving Paradigm for Stem Cell Biology. Cell *132*, 631. 10.1016/j.cell.2008.01.025.

123. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589. 10.1016/j.molcel.2010.05.004.

124. van Oevelen, C., Collombet, S., Vicent, G., Hoogenkamp, M., Lepoivre, C., Badeaux, A., Bussmann, L., Sardina, J.L., Thieffry, D., Beato, M., et al. (2015). C/EBPα Activates Pre-existing and De Novo Macrophage Enhancers during Induced Pre-B Cell Transdifferentiation and Myelopoiesis. Stem Cell Rep. *5*, 232–247. 10.1016/j.stemcr.2015.06.007.

125. Lichtinger, M., Ingram, R., Hannah, R., Müller, D., Clarke, D., Assi, S.A., Lie-A-Ling, M., Noailles, L., Vijayabaskar, M.S., Wu, M., et al. (2012). RUNX1 reshapes the epigenetic landscape at the onset of haematopoiesis. EMBO J. *31*, 4318–4333. 10.1038/emboj.2012.275.

126. de Bruijn, M., and Dzierzak, E. (2017). Runx transcription factors in the development and function of the definitive hematopoietic system. Blood *129*, 2061–2069. 10.1182/blood-2016-12-689109.

127. Johnson, J.L., Georgakilas, G., Petrovic, J., Kurachi, M., Cai, S., Harly, C., Pear, W.S., Bhandoola, A., Wherry, E.J., and Vahedi, G. (2018). Lineage-Determining Transcription Factor TCF-1 Initiates the Epigenetic Identity of T Cells. Immunity *48*, 243-257.e10. 10.1016/j.immuni.2018.01.012.

128. Magor, G.W., Gillinder, K.R., Huang, S., Ilsley, M.D., Bell, C., and Perkins, A.C. (2022). KLF1 Acts As a Pioneer Transcription Factor Via SMARCA4 to Open Chromatin and Facilitate Redeployment of an Enhancer Complex Containing GATA1 and SCL. Blood *140*, 696–697. 10.1182/blood-2022-157901.

129. Buck, M.J., and Lieb, J.D. (2006). A chromatin-mediated mechanism for specification of conditional transcription factor targets. Nat. Genet. *38*, 1446–1451. 10.1038/ng1917.

130. Kouzarides, T. (2007). Chromatin modifications and their function. Cell *128*, 693–705. 10.1016/j.cell.2007.02.005.

131. Marmorstein, R., and Trievel, R.C. (2009). Histone modifying enzymes: structures, mechanisms, and specificities. Biochim. Biophys. Acta *1789*, 58–68. 10.1016/j.bbagrm.2008.07.009.

132. Gourisankar, S., Krokhotin, A., Wenderski, W., and Crabtree, G.R. (2023). Context-specific functions of chromatin remodellers in development and disease. Nat. Rev. Genet. 10.1038/s41576-023-00666-x.

133. Bao, Y., and Shen, X. (2007). INO80 subfamily of chromatin remodeling complexes. Mutat. Res. *618*, 18–29. 10.1016/j.mrfmmm.2006.10.006.

134. Bartholomew, B. (2014). ISWI chromatin remodeling: one primary actor or a coordinated effort? Curr. Opin. Struct. Biol. *24*, 150–155. 10.1016/j.sbi.2014.01.010.

135. Euskirchen, G., Auerbach, R.K., and Snyder, M. (2012). SWI/SNF chromatin-remodeling factors: multiscale analyses and diverse functions. J. Biol. Chem. *287*, 30897–30905. 10.1074/jbc.R111.309302.

136. Alver, B.H., Kim, K.H., Lu, P., Wang, X., Manchester, H.E., Wang, W., Haswell, J.R., Park, P.J., and Roberts, C.W.M. (2017). The SWI/SNF chromatin remodelling complex is required for maintenance of lineage specific enhancers. Nat. Commun. *8*, 14648. 10.1038/ncomms14648.

137. Mittal, P., and Roberts, C.W.M. (2020). The SWI/SNF complex in cancer — biology, biomarkers and therapy. Nat. Rev. Clin. Oncol. *17*, 435–448. 10.1038/s41571-020-0357-3.

138. Mashtalir, N., D'Avino, A.R., Michel, B.C., Luo, J., Pan, J., Otto, J.E., Zullow, H.J., McKenzie, Z.M., Kubiak, R.L., St Pierre, R., et al. (2018). Modular Organization and Assembly of SWI/SNF Family Chromatin Remodeling Complexes. Cell *175*, 1272-1288.e20. 10.1016/j.cell.2018.09.032.

139. Hodges, C., Kirkland, J.G., and Crabtree, G.R. (2016). The Many Roles of BAF (mSWI/SNF) and PBAF Complexes in Cancer. Cold Spring Harb. Perspect. Med. *6*, a026930. 10.1101/cshperspect.a026930.

140. Ogiwara, H., Ui, A., Otsuka, A., Satoh, H., Yokomi, I., Nakajima, S., Yasui, A., Yokota, J., and Kohno, T. (2011). Histone acetylation by CBP and p300 at double-strand break sites facilitates SWI/SNF chromatin remodeling and the recruitment of non-homologous end joining factors. Oncogene *30*, 2135–2146. 10.1038/onc.2010.592.

141. Brownlee, P.M., Meisenberg, C., and Downs, J.A. (2015). The SWI/SNF chromatin remodelling complex: Its role in maintaining genome stability and preventing tumourigenesis. DNA Repair *32*, 127–133. 10.1016/j.dnarep.2015.04.023.

142. Qi, W., Wang, R., Chen, H., Wang, X., Xiao, T., Boldogh, I., Ba, X., Han, L., and Zeng, X. (2015). BRG1 promotes the repair of DNA double-strand breaks by facilitating the replacement of RPA with RAD51. J. Cell Sci. *128*, 317–330. 10.1242/jcs.159103.

143. Watanabe, R., Ui, A., Kanno, S.-I., Ogiwara, H., Nagase, T., Kohno, T., and Yasui, A. (2014). SWI/SNF factors required for cellular resistance to DNA damage include ARID1A and ARID1B and show interdependent protein stability. Cancer Res. *74*, 2465–2475. 10.1158/0008-5472.CAN-13-3608.

144. Chen, Y., Zhang, H., Xu, Z., Tang, H., Geng, A., Cai, B., Su, T., Shi, J., Jiang, C., Tian, X., et al. (2019). A PARP1-BRG1-SIRT1 axis promotes HR repair by reducing nucleosome density at DNA damage sites. Nucleic Acids Res. *47*, 8563–8580. 10.1093/nar/gkz592.

145. Park, J.-H., Park, E.-J., Lee, H.-S., Kim, S.J., Hur, S.-K., Imbalzano, A.N., and Kwon, J. (2006). Mammalian SWI/SNF complexes facilitate DNA double-strand break repair by promoting gamma-H2AX induction. EMBO J. *25*, 3986–3997. 10.1038/sj.emboj.7601291.

146. Shain, A.H., and Pollack, J.R. (2013). The spectrum of SWI/SNF mutations, ubiquitous in human cancers. PloS One *8*, e55119. 10.1371/journal.pone.0055119.

147. Versteege, I., Sévenet, N., Lange, J., Rousseau-Merck, M.F., Ambros, P., Handgretinger, R., Aurias, A., and Delattre, O. (1998). Truncating mutations of hSNF5/INI1 in aggressive paediatric cancer. Nature *394*, 203–206. 10.1038/28212.

148. Roberts, C.W.M., Leroux, M.M., Fleming, M.D., and Orkin, S.H. (2002). Highly penetrant, rapid tumorigenesis through conditional inversion of the tumor suppressor gene Snf5. Cancer Cell *2*, 415–425. 10.1016/s1535-6108(02)00185-x.

149. Varela, I., Tarpey, P., Raine, K., Huang, D., Ong, C.K., Stephens, P., Davies, H., Jones, D., Lin, M.-L., Teague, J., et al. (2011). Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. Nature *469*, 539–542. 10.1038/nature09639.

150. Wiegand, K.C., Shah, S.P., Al-Agha, O.M., Zhao, Y., Tse, K., Zeng, T., Senz, J., McConechy, M.K., Anglesio, M.S., Kalloger, S.E., et al. (2010). ARID1A mutations in endometriosis-associated ovarian carcinomas. N. Engl. J. Med. *363*, 1532–1543. 10.1056/NEJMoa1008433.

151. Jones, S., Wang, T.-L., Shih, I.-M., Mao, T.-L., Nakayama, K., Roden, R., Glas, R., Slamon, D., Diaz, L.A., Vogelstein, B., et al. (2010). Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. Science *330*, 228–231. 10.1126/science.1196333.

152. Kadoch, C., Hargreaves, D.C., Hodges, C., Elias, L., Ho, L., Ranish, J., and Crabtree, G.R. (2013). Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. Nat. Genet. *45*, 592–601. 10.1038/ng.2628.

153. Centore, R.C., Sandoval, G.J., Soares, L.M.M., Kadoch, C., and Chan, H.M. (2020). Mammalian SWI/SNF Chromatin Remodeling Complexes: Emerging Mechanisms and Therapeutic Strategies. Trends Genet. TIG *36*, 936–950. 10.1016/j.tig.2020.07.011.

154. McDonald, E.R., de Weck, A., Schlabach, M.R., Billy, E., Mavrakis, K.J., Hoffman, G.R., Belur, D., Castelletti, D., Frias, E., Gampa, K., et al. (2017). Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening. Cell *170*, 577-592.e10. 10.1016/j.cell.2017.07.005.

155. Helming, K.C., Wang, X., Wilson, B.G., Vazquez, F., Haswell, J.R., Manchester, H.E., Kim, Y., Kryukov, G.V., Ghandi, M., Aguirre, A.J., et al. (2014). ARID1B is a specific vulnerability in ARID1A-mutant cancers. Nat. Med. *20*, 251–254. 10.1038/nm.3480.

156. Minderjahn, J., Schmidt, A., Fuchs, A., Schill, R., Raithel, J., Babina, M., Schmidl, C., Gebhard, C., Schmidhofer, S., Mendes, K., et al. (2020). Mechanisms governing the pioneering and redistribution capabilities of the non-classical pioneer PU.1. Nat. Commun. *11*, 402. 10.1038/s41467-019-13960-2.

157. Chambers, C., Cermakova, K., Chan, Y.S., Kurtz, K., Wohlan, K., Lewis, A.H., Wang, C., Pham, A., Dejmek, M., Sala, M., et al. (2023). SWI/SNF Blockade Disrupts PU.1-Directed Enhancer Programs in Normal Hematopoietic Cells and Acute Myeloid Leukemia. Cancer Res., OF1–OF14. 10.1158/0008-5472.CAN-22-2129.

158. Bakshi, R., Hassan, M.Q., Pratap, J., Lian, J.B., Montecino, M.A., van Wijnen, A.J., Stein, J.L., Imbalzano, A.N., and Stein, G.S. (2010). The human SWI/SNF complex associates with RUNX1 to control transcription of hematopoietic target genes. J. Cell. Physiol. *225*, 569–576. 10.1002/jcp.22240.

159. Müller, C., Calkhoven, C.F., Sha, X., and Leutz, A. (2004). The CCAAT enhancer-binding protein alpha (C/EBPalpha) requires a SWI/SNF complex for proliferation arrest. J. Biol. Chem. *279*, 7353–7358. 10.1074/jbc.M312709200.

160. Kowenz-Leutz, E., and Leutz, A. (1999). A C/EBP beta isoform recruits the SWI/SNF complex to activate myeloid genes. Mol. Cell *4*, 735–743. 10.1016/s1097-2765(00)80384-6.

161. Bultman, S.J., Gebuhr, T.C., and Magnuson, T. (2005). A Brg1 mutation that uncouples ATPase activity from chromatin remodeling reveals an essential role for SWI/SNF-related complexes in beta-globin expression and erythroid development. Genes Dev. *19*, 2849–2861. 10.1101/gad.1364105.

162. Kim, S.-I., Bultman, S.J., Kiefer, C.M., Dean, A., and Bresnick, E.H. (2009). BRG1 requirement for long-range interaction of a locus control region with a downstream promoter. Proc. Natl. Acad. Sci. U. S. A. *106*, 2259–2264. 10.1073/pnas.0806420106.

163. Guo, X., Zhao, Y., Kim, J., and Dean, A. (2022). Hemogen/BRG1 cooperativity modulates promoter and enhancer activation during erythropoiesis. Blood *139*, 3532–3545. 10.1182/blood.2021014308.

164. Armstrong, J.A., Bieker, J.J., and Emerson, B.M. (1998). A SWI/SNF-related chromatin remodeling complex, E-RC1, is required for tissue-specific transcriptional regulation by EKLF in vitro. Cell *95*, 93–104. 10.1016/s0092-8674(00)81785-7.

165. Kim, S.-I., Bresnick, E.H., and Bultman, S.J. (2009). BRG1 directly regulates nucleosome structure and chromatin looping of the alpha globin locus to activate transcription. Nucleic Acids Res. *37*, 6019–6027. 10.1093/nar/gkp677.

166. Grueber, W.B., Jan, L.Y., and Jan, Y.N. (2003). Different levels of the homeodomain protein cut regulate distinct dendrite branching patterns of Drosophila multidendritic neurons. Cell *112*, 805–818. 10.1016/s0092-8674(03)00160-0.

167. Cubelos, B., Sebastián-Serrano, A., Beccari, L., Calcagnotto, M.E., Cisneros, E., Kim, S., Dopazo, A., Alvarez-Dolado, M., Redondo, J.M., Bovolenta, P., et al. (2010). Cux1 and Cux2 regulate dendritic branching, spine morphology and synapses of the upper layer neurons of the cortex. Neuron *66*, 523–535. 10.1016/j.neuron.2010.04.038.

168. Zhai, Z., Ha, N., Papagiannouli, F., Hamacher-Brady, A., Brady, N., Sorge, S., Bezdan, D., and Lohmann, I. (2012). Antagonistic regulation of apoptosis and differentiation by the Cut transcription factor represents a tumor-suppressing mechanism in Drosophila. PLoS Genet. *8*, e1002582. 10.1371/journal.pgen.1002582.

169. Arthur, R.K., An, N., Khan, S., and McNerney, M.E. (2017). The haploinsufficient tumor suppressor, CUX1, acts as an analog transcriptional regulator that controls target genes through distal enhancers that loop to target promoters. Nucleic Acids Res. *45*, 6350–6361. 10.1093/nar/gkx218.

170. Last, T.J., van Wijnen, A.J., de Ridder, M.C., Stein, G.S., and Stein, J.L. (1999). The homeodomain transcription factor CDP/cut interacts with the cell cycle regulatory element of histone H4 genes packaged into nucleosomes. Mol. Biol. Rep. *26*, 185–194. 10.1023/a:1007058123699.

171. Zhu, F., Farnung, L., Kaasinen, E., Sahu, B., Yin, Y., Wei, B., Dodonova, S.O., Nitta, K.R., Morgunova, E., Taipale, M., et al. (2018). The interaction landscape between transcription factors and the nucleosome. Nature *562*, 76–81. 10.1038/s41586-018-0549-5.

172. Dzierzak, E., and Philipsen, S. (2013). Erythropoiesis: development and differentiation. Cold Spring Harb. Perspect. Med. *3*, a011601. 10.1101/cshperspect.a011601.

173. Palis, J. (2014). Primitive and definitive erythropoiesis in mammals. Front. Physiol. *5*, 3. 10.3389/fphys.2014.00003.

174. Wang, X., and Thein, S.L. (2018). Switching from fetal to adult hemoglobin. Nat. Genet. *50*, 478–480. 10.1038/s41588-018-0094-z.

175. Valent, P., Büsche, G., Theurl, I., Uras, I.Z., Germing, U., Stauder, R., Sotlar, K., Füreder, W., Bettelheim, P., Pfeilstöcker, M., et al. (2018). Normal and pathological erythropoiesis in adults: from gene regulation to targeted treatment concepts. Haematologica *103*, 1593–1603. 10.3324/haematol.2018.192518.

176. Zivot, A., Lipton, J.M., Narla, A., and Blanc, L. (2018). Erythropoiesis: insights into pathophysiology and treatments in 2017. Mol. Med. Camb. Mass *24*, 11. 10.1186/s10020-018-0011-z.

177. Kim, S.-I., and Bresnick, E.H. (2007). Transcriptional control of erythropoiesis: emerging mechanisms and principles. Oncogene *26*, 6777–6794. 10.1038/sj.onc.1210761.

178. Hattangadi, S.M., Wong, P., Zhang, L., Flygare, J., and Lodish, H.F. (2011). From stem cell to red cell: regulation of erythropoiesis at multiple levels by multiple proteins, RNAs, and chromatin modifications. Blood *118*, 6258–6268. 10.1182/blood-2011-07-356006.

179. Koury, M.J., Sawyer, S.T., and Brandt, S.J. (2002). New insights into erythropoiesis. Curr. Opin. Hematol. *9*, 93–100. 10.1097/00062752-200203000-00002.

180. Ferreira, R., Ohneda, K., Yamamoto, M., and Philipsen, S. (2005). GATA1 Function, a Paradigm for Transcription Factors in Hematopoiesis. Mol. Cell. Biol. *25*, 1215–1227. 10.1128/MCB.25.4.1215-1227.2005.

181. Simon, M.C., Pevny, L., Wiles, M.V., Keller, G., Costantini, F., and Orkin, S.H. (1992). Rescue of erythroid development in gene targeted GATA-1- mouse embryonic stem cells. Nat. Genet. *1*, 92–98. 10.1038/ng0592-92.

182. Haase, V.H. (2010). Hypoxic regulation of erythropoiesis and iron metabolism. Am. J. Physiol. Renal Physiol. *299*, F1-13. 10.1152/ajprenal.00174.2010.

183. Bhoopalan, S.V., Huang, L.J.-S., and Weiss, M.J. (2020). Erythropoietin regulation of red blood cell production: from bench to bedside and back. F1000Research *9*, F1000 Faculty Rev-1153. 10.12688/f1000research.26648.1.

184. Jafari, M., Ghadami, E., Dadkhah, T., and Akhavan-Niaki, H. (2019). PI3k/AKT signaling pathway: Erythropoiesis and beyond. J. Cell. Physiol. *234*, 2373–2385. 10.1002/jcp.27262.

185. Watts, D., Gaete, D., Rodriguez, D., Hoogewijs, D., Rauner, M., Sormendi, S., and Wielockx, B. (2020). Hypoxia Pathway Proteins are Master Regulators of Erythropoiesis. Int. J. Mol. Sci. *21*, 8131. 10.3390/ijms21218131.

186. Hasserjian, R.P., Howard, J., Wood, A., Henry, K., and Bain, B. (2001). Acute erythremic myelosis (true erythroleukaemia): a variant of AML FAB-M6. J. Clin. Pathol. *54*, 205–209. 10.1136/jcp.54.3.205.

187. Yu, M., Riva, L., Xie, H., Schindler, Y., Moran, T.B., Cheng, Y., Yu, D., Hardison, R., Weiss, M.J., Orkin, S.H., et al. (2009). Insights into GATA-1-mediated gene activation versus repression via genome-wide chromatin occupancy analysis. Mol. Cell *36*, 682–695. 10.1016/j.molcel.2009.11.002.

188. Tsai, S.F., Martin, D.I., Zon, L.I., D'Andrea, A.D., Wong, G.G., and Orkin, S.H. (1989). Cloning of cDNA for the major DNA-binding protein of the erythroid lineage through expression in mammalian cells. Nature *339*, 446–451. 10.1038/339446a0.

189. Evans, T., and Felsenfeld, G. (1989). The erythroid-specific transcription factor Eryf1: a new finger protein. Cell *58*, 877–885. 10.1016/0092-8674(89)90940-9.

190. Martin, D.I., Zon, L.I., Mutter, G., and Orkin, S.H. (1990). Expression of an erythroid transcription factor in megakaryocytic and mast cell lineages. Nature *344*, 444–447. 10.1038/344444a0.

191. Gutiérrez, L., Nikolic, T., van Dijk, T.B., Hammad, H., Vos, N., Willart, M., Grosveld, F., Philipsen, S., and Lambrecht, B.N. (2007). Gata1 regulates dendritic-cell development and survival. Blood *110*, 1933–1941. 10.1182/blood-2006-09-048322.

192. Leonard, M., Brice, M., Engel, J.D., and Papayannopoulou, T. (1993). Dynamics of GATA transcription factor expression during erythroid differentiation. Blood *82*, 1071–1079.

193. Vyas, P., Ault, K., Jackson, C.W., Orkin, S.H., and Shivdasani, R.A. (1999). Consequences of GATA-1 deficiency in megakaryocytes and platelets. Blood *93*, 2867–2875.

194. Yu, C., Cantor, A.B., Yang, H., Browne, C., Wells, R.A., Fujiwara, Y., and Orkin, S.H. (2002). Targeted deletion of a high-affinity GATA-binding site in the GATA-1 promoter leads to selective loss of the eosinophil lineage in vivo. J. Exp. Med. *195*, 1387–1395. 10.1084/jem.20020656.

195. Nei, Y., Obata-Ninomiya, K., Tsutsui, H., Ishiwata, K., Miyasaka, M., Matsumoto, K., Nakae, S., Kanuka, H., Inase, N., and Karasuyama, H. (2013). GATA-1 regulates the generation and function of basophils. Proc. Natl. Acad. Sci. U. S. A. *110*, 18620–18625. 10.1073/pnas.1311668110.

196. Bresnick, E.H., Katsumura, K.R., Lee, H.-Y., Johnson, K.D., and Perkins, A.S. (2012). Master regulatory GATA transcription factors: mechanistic principles and emerging links to hematologic malignancies. Nucleic Acids Res. *40*, 5819–5831. 10.1093/nar/gks281.

197. Katsumura, K.R., Bresnick, E.H., and GATA Factor Mechanisms Group (2017). The GATA factor revolution in hematology. Blood *129*, 2092–2102. 10.1182/blood-2016-09-687871.

198. Molkentin, J.D. (2000). The zinc finger-containing transcription factors GATA-4, -5, and -6. Ubiquitously expressed regulators of tissue-specific gene expression. J. Biol. Chem. *275*, 38949–38952. 10.1074/jbc.R000029200.

199. Charron, F., and Nemer, M. (1999). GATA transcription factors and cardiac development. Semin. Cell Dev. Biol. *10*, 85–91. 10.1006/scdb.1998.0281.

200. Vicente, C., Conchillo, A., García-Sánchez, M.A., and Odero, M.D. (2012). The role of the GATA2 transcription factor in normal and malignant hematopoiesis. Crit. Rev. Oncol. Hematol. *82*, 1–17. 10.1016/j.critrevonc.2011.04.007.

201. Fujiwara, Y., Chang, A.N., Williams, A.M., and Orkin, S.H. (2004). Functional overlap of GATA-1 and GATA-2 in primitive hematopoietic development. Blood *103*, 583–585. 10.1182/blood-2003-08-2870.

202. Grass, J.A., Boyer, M.E., Pal, S., Wu, J., Weiss, M.J., and Bresnick, E.H. (2003). GATA-1-dependent transcriptional repression of GATA-2 via disruption of positive autoregulation and domain-wide chromatin remodeling. Proc. Natl. Acad. Sci. U. S. A. *100*, 8811–8816. 10.1073/pnas.1432147100.

203. Huang, J., Liu, X., Li, D., Shao, Z., Cao, H., Zhang, Y., Trompouki, E., Bowman, T.V., Zon, L.I., Yuan, G.-C., et al. (2016). Dynamic Control of Enhancer Repertoires Drives Lineage and Stage-Specific Transcription during Hematopoiesis. Dev. Cell *36*, 9–23. 10.1016/j.devcel.2015.12.014.

204. Han, G.C., Vinayachandran, V., Bataille, A.R., Park, B., Chan-Salis, K.Y., Keller, C.A., Long, M., Mahony, S., Hardison, R.C., and Pugh, B.F. (2016). Genome-Wide Organization of GATA1 and TAL1 Determined at High Resolution. Mol. Cell. Biol. *36*, 157–172. 10.1128/MCB.00806-15.

205. Tallack, M.R., Magor, G.W., Dartigues, B., Sun, L., Huang, S., Fittock, J.M., Fry, S.V., Glazov, E.A., Bailey, T.L., and Perkins, A.C. (2012). Novel roles for KLF1 in erythropoiesis revealed by mRNA-seq. Genome Res. *22*, 2385–2398. 10.1101/gr.135707.111.

206. Kim, S.-I., Bultman, S.J., Jing, H., Blobel, G.A., and Bresnick, E.H. (2007). Dissecting molecular steps in chromatin domain activation during hematopoietic differentiation. Mol. Cell. Biol. *27*, 4551–4565. 10.1128/MCB.00235-07.

207. Zhang, P., Zhang, X., Iwama, A., Yu, C., Smith, K.A., Mueller, B.U., Narravula, S., Torbett, B.E., Orkin, S.H., and Tenen, D.G. (2000). PU.1 inhibits GATA-1 function and erythroid differentiation by blocking GATA-1 DNA binding. Blood *96*, 2641–2648.

208. Rhodes, J., Hagen, A., Hsu, K., Deng, M., Liu, T.X., Look, A.T., and Kanki, J.P. (2005). Interplay of pu.1 and gata1 determines myelo-erythroid progenitor cell fate in zebrafish. Dev. Cell *8*, 97–108. 10.1016/j.devcel.2004.11.014.

209. Gundry, M.C., Brunetti, L., Lin, A., Mayle, A.E., Kitano, A., Wagner, D., Hsu, J.I., Hoegenauer, K.A., Rooney, C.M., Goodell, M.A., et al. (2016). Highly Efficient Genome Editing of Murine and Human Hematopoietic Progenitor Cells by CRISPR/Cas9. Cell Rep. *17*, 1453–1461. 10.1016/j.celrep.2016.09.092.

210. Castaño, J., Bueno, C., Jiménez-Delgado, S., Roca-Ho, H., Fraga, M.F., Fernandez, A.F., Nakanishi, M., Torres-Ruiz, R., Rodríguez-Perales, S., and Menéndez, P. (2017). Generation and characterization of a human iPSC cell line expressing inducible Cas9 in the "safe harbor" AAVS1 locus. Stem Cell Res. *21*, 137–140. 10.1016/j.scr.2017.04.011.

211. Morotti, A., Panuzzo, C., Crivellaro, S., Carrà, G., Torti, D., Guerrasio, A., and Saglio, G. (2015). The Role of PTEN in Myeloid Malignancies. Hematol. Rep. *7*, 5844. 10.4081/hr.2015.6027.

212. Welch, J.J., Watts, J.A., Vakoc, C.R., Yao, Y., Wang, H., Hardison, R.C., Blobel, G.A., Chodosh, L.A., and Weiss, M.J. (2004). Global regulation of erythroid gene expression by transcription factor GATA-1. Blood *104*, 3136–3147. 10.1182/blood-2004-04-1603.

213. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., and Forbes, S.A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat. Rev. Cancer *18*, 696–705. 10.1038/s41568-018-0060-1.

214. Baeten, J.T., Liu, W., Preddy, I.C., Zhou, N., and McNerney, M.E. (2022). CRISPR screening in human hematopoietic stem and progenitor cells reveals an enrichment for tumor suppressor genes within chromosome 7 commonly deleted regions. Leukemia *36*, 1421–1425. 10.1038/s41375-021-01491-z.

215. Brunetti, L., Gundry, M.C., Kitano, A., Nakada, D., and Goodell, M.A. (2018). Highly Efficient Gene Disruption of Murine and Human Hematopoietic Progenitor Cells by CRISPR/Cas9. J. Vis. Exp. JoVE, 57278. 10.3791/57278.

216. Brinkman, E.K., Kousholt, A.N., Harmsen, T., Leemans, C., Chen, T., Jonkers, J., and Steensel, B. van (2017). Easy quantification of template-directed CRISPR/Cas9 editing. Preprint at bioRxiv, 10.1101/218156 10.1101/218156.

217. Skene, P.J., and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. eLife *6*, e21856. 10.7554/eLife.21856.

218. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218. 10.1038/nmeth.2688.

219. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12. 10.14806/ej.17.1.200.

220. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. *25*, 1754–1760. 10.1093/bioinformatics/btp324.

221. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137. 10.1186/gb-2008-9-9-r137.

222. Li, Q., Brown, J.B., Huang, H., and Bickel, P.J. (2011). Measuring reproducibility of high-throughput experiments. Ann. Appl. Stat. *5*, 1752–1779. 10.1214/11-AOAS466.

223. Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci. Rep. *9*, 9354. 10.1038/s41598-019-45839-z.

224. Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. *42*, W187–W191. 10.1093/nar/gku365.

225. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26. 10.1038/nbt.1754.

226. McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495–501. 10.1038/nbt.1630.

227. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. 10.1093/bioinformatics/btq033.

228. Zhu, L.J., Gazin, C., Lawson, N.D., Pagès, H., Lin, S.M., Lapointe, D.S., and Green, M.R. (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics *11*, 237. 10.1186/1471-2105-11-237.

229. Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. Bioinforma. Oxf. Engl. *27*, 1696–1697. 10.1093/bioinformatics/btr189.

230. Tl, B., J, J., Ce, G., and Ws, N. (2015). The MEME Suite. Nucleic Acids Res. *43*. 10.1093/nar/gkv416.

231. McLeay, R.C., and Bailey, T.L. (2010). Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics *11*, 165. 10.1186/1471-2105-11-165.

232. Lun, A.T.L., and Smyth, G.K. (2016). csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. Nucleic Acids Res. *44*, e45. 10.1093/nar/gkv1191.

233. Zhang, Y., and Hardison, R.C. (2017). Accurate and reproducible functional maps in 127 human cell types via 2D genome segmentation. Nucleic Acids Res. *45*, 9823–9836. 10.1093/nar/gkx659.

234. Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol. *28*, 817–825. 10.1038/nbt.1662.

235. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49. 10.1038/nature09906.

236. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330. 10.1038/nature14248.

237. Zhang, X., Jeong, M., Huang, X., Wang, X.Q., Wang, X., Zhou, W., Shamim, M.S., Gore, H., Himadewi, P., Liu, Y., et al. (2020). Large DNA Methylation Nadirs Anchor Chromatin Loops Maintaining Hematopoietic Stem Cell Identity. Mol. Cell *78*, 506-521.e6. 10.1016/j.molcel.2020.04.018.

238. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., and Klein, A.M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science *367*, eaaw3381. 10.1126/science.aaw3381.

239. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573-3587.e29. 10.1016/j.cell.2021.04.048.

240. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

241. Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191. 10.1038/nprot.2009.97.

242. Novershtern, N., Subramanian, A., Lawton, L.N., Mak, R.H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., Shay, T., et al. (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell *144*, 296–309. 10.1016/j.cell.2011.01.004.

243. Kim, K.M., Mura-Meszaros, A., Tollot, M., Krishnan, M.S., Gründl, M., Neubert, L., Groth, M., Rodriguez-Fraticelli, A., Svendsen, A.F., Campaner, S., et al. (2022). Taz protects hematopoietic stem cells from an aging-dependent decrease in PU.1 activity. Nat. Commun. *13*, 5187. 10.1038/s41467-022-32970-1.

244. Cai, X., Gao, L., Teng, L., Ge, J., Oo, Z.M., Kumar, A.R., Gilliland, D.G., Mason, P.J., Tan, K., and Speck, N.A. (2015). Runx1 Deficiency Decreases Ribosome Biogenesis and Confers Stress Resistance to Hematopoietic Stem and Progenitor Cells. Cell Stem Cell *17*, 165–177. 10.1016/j.stem.2015.06.002.

245. Wu, J.Q., Seay, M., Schulz, V.P., Hariharan, M., Tuck, D., Lian, J., Du, J., Shi, M., Ye, Z., Gerstein, M., et al. (2012). Tcf7 Is an Important Regulator of the Switch of Self-Renewal and Differentiation in a Multipotential Hematopoietic Cell Line. PLOS Genet. *8*, e1002565. 10.1371/journal.pgen.1002565.

246. Hu, H., Miao, Y.-R., Jia, L.-H., Yu, Q.-Y., Zhang, Q., and Guo, A.-Y. (2019). AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. Nucleic Acids Res. *47*, D33–D38. 10.1093/nar/gky822.

247. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. 10.1093/bioinformatics/bts635.

248. Zappia, L., and Oshlack, A. (2018). Clustering trees: a visualization for evaluating clusterings at multiple resolutions. GigaScience *7*, giy083. 10.1093/gigascience/giy083.

249. Aran, D., Looney, A.P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R.P., Wolters, P.J., Abate, A.R., et al. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. *20*, 163–172. 10.1038/s41590-018-0276-y.

250. McGinnis, C.S., Murrow, L.M., and Gartner, Z.J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst. *8*, 329-337.e4. 10.1016/j.cels.2019.03.003.

251. Van den Berge, K., Roux de Bézieux, H., Street, K., Saelens, W., Cannoodt, R., Saeys, Y., Dudoit, S., and Clement, L. (2020). Trajectory-based differential expression analysis for single-cell sequencing data. Nat. Commun. *11*, 1201. 10.1038/s41467-020-14766-3.

252. Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics *19*, 477. 10.1186/s12864-018-4772-0.

253. Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. Nat. Biotechnol. *37*, 547–554. 10.1038/s41587-019-0071-9.

254. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinforma. Oxf. Engl. *26*, 139–140. 10.1093/bioinformatics/btp616.

255. Inaba, T., Honda, H., and Matsui, H. (2018). The enigma of monosomy 7. Blood *131*, 2891–2898. 10.1182/blood-2017-12-822262.

256. Asou, H., Matsui, H., Ozaki, Y., Nagamachi, A., Nakamura, M., Aki, D., and Inaba, T. (2009). Identification of a common microdeletion cluster in 7q21.3 subband among patients with myeloid leukemia and myelodysplastic syndrome. Biochem. Biophys. Res. Commun. *383*, 245–251. 10.1016/j.bbrc.2009.04.004.

257. Stanford, W.L., Cohn, J.B., and Cordes, S.P. (2001). Gene-trap mutagenesis: past, present and beyond. Nat. Rev. Genet. *2*, 756–768. 10.1038/35093548.

258. Mohr, S.E., and Perrimon, N. (2012). RNAi screening: new approaches, understandings, and organisms. Wiley Interdiscip. Rev. RNA *3*, 145–158. 10.1002/wrna.110.

259. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science *337*, 816–821. 10.1126/science.1225829.

260. Bikard, D., Jiang, W., Samai, P., Hochschild, A., Zhang, F., and Marraffini, L.A. (2013). Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. Nucleic Acids Res. *41*, 7429–7437. 10.1093/nar/gkt520.

261. Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell *152*, 1173–1183. 10.1016/j.cell.2013.02.022.

262. Bock, C., Datlinger, P., Chardon, F., Coelho, M.A., Dong, M.B., Lawson, K.A., Lu, T., Maroc, L., Norman, T.M., Song, B., et al. (2022). High-content CRISPR screening. Nat. Rev. Methods Primer *2*, 9. 10.1038/s43586-022-00098-7.

263. Blomen, V.A., Májek, P., Jae, L.T., Bigenzahn, J.W., Nieuwenhuis, J., Staring, J., Sacco, R., van Diemen, F.R., Olk, N., Stukalov, A., et al. (2015). Gene essentiality and synthetic lethality in haploid human cells. Science *350*, 1092–1096. 10.1126/science.aac7557.

264. Gilbert, L.A., Horlbeck, M.A., Adamson, B., Villalta, J.E., Chen, Y., Whitehead, E.H., Guimaraes, C., Panning, B., Ploegh, H.L., Bassik, M.C., et al. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell *159*, 647–661. 10.1016/j.cell.2014.09.029.

265. Wang, T., Birsoy, K., Hughes, N.W., Krupczak, K.M., Post, Y., Wei, J.J., Lander, E.S., and Sabatini, D.M. (2015). Identification and characterization of essential genes in the human genome. Science *350*, 1096–1101. 10.1126/science.aac7041.

266. Wang, T., Yu, H., Hughes, N.W., Liu, B., Kendirli, A., Klein, K., Chen, W.W., Lander, E.S., and Sabatini, D.M. (2017). Gene Essentiality Profiling Reveals Gene Networks and Synthetic Lethal Interactions with Oncogenic Ras. Cell *168*, 890-903.e15. 10.1016/j.cell.2017.01.013.

267. Wallace, J., Hu, R., Mosbruger, T.L., Dahlem, T.J., Stephens, W.Z., Rao, D.S., Round, J.L., and O'Connell, R.M. (2016). Genome-Wide CRISPR-Cas9 Screen Identifies MicroRNAs That Regulate Myeloid Leukemia Cell Growth. PloS One *11*, e0153689. 10.1371/journal.pone.0153689.

268. Horlbeck, M.A., Gilbert, L.A., Villalta, J.E., Adamson, B., Pak, R.A., Chen, Y., Fields, A.P., Park, C.Y., Corn, J.E., Kampmann, M., et al. (2016). Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. eLife *5*, e19760. 10.7554/eLife.19760.

269. Sack, L.M., Davoli, T., Li, M.Z., Li, Y., Xu, Q., Naxerova, K., Wooten, E.C., Bernardi, R.J., Martin, T.D., Chen, T., et al. (2018). Profound Tissue Specificity in Proliferation Control Underlies Cancer Drivers and Aneuploidy Patterns. Cell *173*, 499-514.e23. 10.1016/j.cell.2018.02.037.

270. Sanson, K.R., Hanna, R.E., Hegde, M., Donovan, K.F., Strand, C., Sullender, M.E., Vaimberg, E.W., Goodale, A., Root, D.E., Piccioni, F., et al. (2018). Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. Nat. Commun. *9*, 5416. 10.1038/s41467-018-07901-8.

271. Bakke, J., Wright, W.C., Zamora, A.E., Oladimeji, P., Crawford, J.C., Brewer, C.T., Autry, R.J., Evans, W.E., Thomas, P.G., and Chen, T. (2019). Genome-wide CRISPR screen reveals PSMA6 to be an essential gene in pancreatic cancer cells. BMC Cancer *19*, 253. 10.1186/s12885-019-5455-1.

272. Sarker, I.H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput. Sci. *2*, 160. 10.1007/s42979-021-00592-x.

273. Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. J. Stat. Softw. *28*, 1–26. 10.18637/jss.v028.i05.

274. Fawagreh, K., Gaber, M.M., and Elyan, E. (2014). Random forests: from early developments to recent advancements. Syst. Sci. Control Eng. *2*, 602–609. 10.1080/21642583.2014.956265.

275. Jäger, S., Allhorn, A., and Bießmann, F. (2021). A Benchmark for Data Imputation Methods. Front. Big Data *4*.

276. Tisserand, J., Khetchoumian, K., Thibault, C., Dembélé, D., Chambon, P., and Losson, R. (2011). Tripartite motif 24 (Trim24/Tif1α) tumor suppressor protein is a novel negative regulator of interferon (IFN)/signal transducers and activators of transcription (STAT) signaling pathway acting through retinoic acid receptor α (Rarα) inhibition. J. Biol. Chem. *286*, 33369–33379. 10.1074/jbc.M111.225680.

277. Makishima, H., Visconte, V., Sakaguchi, H., Jankowska, A.M., Abu Kar, S., Jerez, A., Przychodzen, B., Bupathi, M., Guinta, K., Afable, M.G., et al. (2012). Mutations in the spliceosome machinery, a novel and ubiquitous pathway in leukemogenesis. Blood *119*, 3203–3210. 10.1182/blood-2011-12-399774.

278. Sherr, C.J. (2004). Principles of tumor suppression. Cell *116*, 235–246. 10.1016/s0092-8674(03)01075-4.

279. Cheng, M.W., Mitra, M., and Coller, H.A. (2023). Pan-cancer landscape of epigenetic factor expression predicts tumor outcome. Commun. Biol. *6*, 1138. 10.1038/s42003-023-05459-w.

280. Sudhakar, M., Rengaswamy, R., and Raman, K. (2022). Multi-Omic Data Improve Prediction of Personalized Tumor Suppressors and Oncogenes. Front. Genet. *13*, 854190. 10.3389/fgene.2022.854190.

281. Bordeleau, F. (2023). Using Machine Learning to Predict TP53 Mutation Status and Aggressiveness of Prostate Cancer from Routine Histology Images. Cancer Res. *83*, 2809–2810. 10.1158/0008-5472.CAN-23-1856.

282. Anandanadarajah, N., Chu, C.H., and Loganantharaj, R. (2021). An integrated deep learning and dynamic programming method for predicting tumor suppressor genes, oncogenes, and fusion from PDB structures. Comput. Biol. Med. *133*, 104323. 10.1016/j.compbiomed.2021.104323.

283. Shen, L., Shi, Q., and Wang, W. (2018). Double agents: genes with both oncogenic and tumor-suppressor functions. Oncogenesis *7*, 25. 10.1038/s41389-018-0034-x.

284. Datta, N., Chakraborty, S., Basu, M., and Ghosh, M.K. (2020). Tumor Suppressors Having Oncogenic Functions: The Double Agents. Cells *10*, 46. 10.3390/cells10010046.

285. Sood, R., Kamikubo, Y., and Liu, P. (2017). Role of RUNX1 in hematological malignancies. Blood *129*, 2070–2082. 10.1182/blood-2016-10-687830.

286. Goyama, S., Schibler, J., Cunningham, L., Zhang, Y., Rao, Y., Nishimoto, N., Nakagawa, M., Olsson, A., Wunderlich, M., Link, K.A., et al. (2013). Transcription factor RUNX1 promotes survival of acute myeloid leukemia cells. J. Clin. Invest. *123*, 3876–3888. 10.1172/JCI68557.

287. Kadam, S., McAlpine, G.S., Phelan, M.L., Kingston, R.E., Jones, K.A., and Emerson, B.M. (2000). Functional selectivity of recombinant mammalian SWI/SNF subunits. Genes Dev. *14*, 2441–2451.

288. Platzer, K., Cogné, B., Hague, J., Marcelis, C.L., Mitter, D., Oberndorff, K., Park, S.-M., Ploos van Amstel, H.K., Simonic, I., van der Smagt, J.J., et al. (2018).

Haploinsufficiency of CUX1 Causes Nonsyndromic Global Developmental Delay With Possible Catch-up Development. Ann. Neurol. *84*, 200–207. 10.1002/ana.25278.

289. Whalen, S., Truty, R.M., and Pollard, K.S. (2016). Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat. Genet. *48*, 488–496. 10.1038/ng.3539.

290. Coqueret, O., Bérubé, G., and Nepveu, A. (1998). The mammalian Cut homeodomain protein functions as a cell-cycle-dependent transcriptional repressor which downmodulates p21WAF1/CIP1/SDI1 in S phase. EMBO J. *17*, 4680–4694. 10.1093/emboj/17.16.4680.

291. Livingston, S., Carlton, C., Sharma, M., Kearns, D., Baybutt, R., and Vanden Heuvel, G.B. (2019). Cux1 regulation of the cyclin kinase inhibitor p27kip1 in polycystic kidney disease is attenuated by HDAC inhibitors. Gene X *2*, 100007. 10.1016/j.gene.2019.100007.

292. Ueda, Y., Su, Y., and Richmond, A. (2007). CCAAT displacement protein regulates nuclear factor-kappa beta-mediated chemokine transcription in melanoma cells. Melanoma Res. *17*, 91–103. 10.1097/CMR.0b013e3280a60888.

293. Truscott, M., Harada, R., Vadnais, C., Robert, F., and Nepveu, A. (2008). p110 CUX1 Cooperates with E2F Transcription Factors in the Transcriptional Activation of Cell Cycle-Regulated Genes. Mol. Cell. Biol. *28*, 3127–3138. 10.1128/MCB.02089-07.

294. Andersson, L.C., Jokinen, M., and Gahmberg, C.G. (1979). Induction of erythroid differentiation in the human leukaemia cell line K562. Nature *278*, 364–365. 10.1038/278364a0.

295. Tabilio, A., Pelicci, P.G., Vinci, G., Mannoni, P., Civin, C.I., Vainchenker, W., Testa, U., Lipinski, M., Rochant, H., and Breton-Gorius, J. (1983). Myeloid and megakaryocytic properties of K-562 cell lines. Cancer Res. *43*, 4569–4574.

296. Green, A.R., Rockman, S., DeLuca, E., and Begley, C.G. (1993). Induced myeloid differentiation of K562 cells with downregulation of erythroid and megakaryocytic transcription factors: a novel experimental model for hemopoietic lineage restriction. Exp. Hematol. *21*, 525–531.

297. Mardinian, K., Adashek, J.J., Botta, G.P., Kato, S., and Kurzrock, R. (2021). SMARCA4: Implications of an altered chromatin-remodeling gene for cancer development and therapy. Mol. Cancer Ther., 10.1158/1535-7163.MCT-21–0433. 10.1158/1535-7163.MCT-21-0433.

298. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. 10.1038/nature11247.

299. Luo, Y., Hitz, B.C., Gabdank, I., Hilton, J.A., Kagda, M.S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., et al. (2020). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. *48*, D882–D889. 10.1093/nar/gkz1062.

300. Elder, G.H. (1998). Genetic Defects in the Porphyrias: Types and Significance. Clin. Dermatol. *16*, 225–233. 10.1016/S0738-081X(97)00202-2.

301. Miller, I.J., and Bieker, J.J. (1993). A novel, erythroid cell-specific murine transcription factor that binds to the CACCC element and is related to the Krüppel family of nuclear proteins. Mol. Cell. Biol. *13*, 2776–2786. 10.1128/mcb.13.5.2776-2786.1993.

302. Nuez, B., Michalovich, D., Bygrave, A., Ploemacher, R., and Grosveld, F. (1995). Defective haematopoiesis in fetal liver resulting from inactivation of the EKLF gene. Nature *375*, 316–318. 10.1038/375316a0.

303. Lyu, Z.-Z., Zhao, B.-B., Koiwai, K., Hirono, I., and Kondo, H. (2016). Identification of endonuclease domain-containing 1 gene in Japanese flounder Paralichthys olivaceus. Fish Shellfish Immunol. *50*, 43–49. 10.1016/j.fsi.2016.01.017.

304. Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., and Lewis, S. (2009). AmiGO: online access to ontology and annotation data. Bioinformatics *25*, 288–289. 10.1093/bioinformatics/btn615.

305. Gene Ontology Consortium (2021). The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. *49*, D325–D334. 10.1093/nar/gkaa1113.

306. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29. 10.1038/75556.

307. Speck, N.A., and Gilliland, D.G. (2002). Core-binding factors in haematopoiesis and leukaemia. Nat. Rev. Cancer *2*, 502–513. 10.1038/nrc840.

308. Nottingham, W.T., Jarratt, A., Burgess, M., Speck, C.L., Cheng, J.-F., Prabhakar, S., Rubin, E.M., Li, P.-S., Sloane-Stanley, J., Kong-a-San, J., et al. (2007). Runx1-mediated hematopoietic stem-cell emergence is controlled by a Gata/Ets/SCL-regulated enhancer. Blood *110*, 4188–4197. 10.1182/blood-2007-07-100883.

309. Anguita, E., Villegas, A., Iborra, F., and Hernández, A. (2010). GFI1B controls its own expression binding to multiple sites. Haematologica *95*, 36–46. 10.3324/haematol.2009.012351.

310. Sato, S., Tomomori-Sato, C., Banks, C.A.S., Sorokina, I., Parmely, T.J., Kong, S.E., Jin, J., Cai, Y., Lane, W.S., Brower, C.S., et al. (2003). Identification of mammalian Mediator subunits with similarities to yeast Mediator subunits Srb5, Srb6, Med11, and Rox3. J. Biol. Chem. *278*, 15123–15127. 10.1074/jbc.C300054200.

311. Ferreira, L.T., Figueiredo, A.C., Orr, B., Lopes, D., and Maiato, H. (2018). Dissecting the role of the tubulin code in mitosis. Methods Cell Biol. *144*, 33–74. 10.1016/bs.mcb.2018.03.040.

312. Wahlestedt, M., Ladopoulos, V., Hidalgo, I., Sanchez Castillo, M., Hannah, R., Säwén, P., Wan, H., Dudenhöffer-Pfeifer, M., Magnusson, M., Norddahl, G.L., et al. (2017).

Critical Modulation of Hematopoietic Lineage Fate by Hepatic Leukemia Factor. Cell Rep. *21*, 2251–2263. 10.1016/j.celrep.2017.10.112.

313. Mancias, J.D., Pontano Vaites, L., Nissim, S., Biancur, D.E., Kim, A.J., Wang, X., Liu, Y., Goessling, W., Kimmelman, A.C., and Harper, J.W. (2015). Ferritinophagy via NCOA4 is required for erythropoiesis and is regulated by iron dependent HERC2-mediated proteolysis. eLife *4*, e10308. 10.7554/eLife.10308.

314. Wang, L., Wang, X., Wang, L., Yousaf, M., Li, J., Zuo, M., Yang, Z., Gou, D., Bao, B., Li, L., et al. (2018). Identification of a new adtrp1-tfpi regulatory axis for the specification of primitive myelopoiesis and definitive hematopoiesis. FASEB J. *32*, 183–194. 10.1096/fj.201700166RR.

315. Thorén, L.A., Liuba, K., Bryder, D., Nygren, J.M., Jensen, C.T., Qian, H., Antonchuk, J., and Jacobsen, S.-E.W. (2008). Kit regulates maintenance of quiescent hematopoietic stem cells. J. Immunol. Baltim. Md 1950 *180*, 2045–2053. 10.4049/jimmunol.180.4.2045.

316. Zaret, K.S., and Mango, S.E. (2016). Pioneer transcription factors, chromatin dynamics, and cell fate control. Curr. Opin. Genet. Dev. *37*, 76–81. 10.1016/j.gde.2015.12.003.

317. Ng, S.Y.-M., Yoshida, T., Zhang, J., and Georgopoulos, K. (2009). Genome-wide lineage-specific transcriptional networks underscore Ikaros-dependent lymphoid priming in hematopoietic stem cells. Immunity *30*, 493–507. 10.1016/j.immuni.2009.01.014.

318. Xiang, G., Keller, C.A., Heuston, E., Giardine, B.M., An, L., Wixom, A.Q., Miller, A., Cockburn, A., Sauria, M.E.G., Weaver, K., et al. (2020). An integrative view of the regulatory and transcriptional landscapes in mouse hematopoiesis. Genome Res. *30*, 472–484. 10.1101/gr.255760.119.

319. Wagner, D.E., and Klein, A.M. (2020). Lineage tracing meets single-cell omics: opportunities and challenges. Nat. Rev. Genet. *21*, 410–427. 10.1038/s41576-020-0223-2.

320. Swift, M.L., Beishline, K., and Azizkhan-Clifford, J. (2021). Sp1-dependent recruitment of the histone acetylase p300 to DSBs facilitates chromatin remodeling and recruitment of the NHEJ repair factor Ku70. DNA Repair *105*, 103171. 10.1016/j.dnarep.2021.103171.

321. Andrades, A., Peinado, P., Alvarez-Perez, J.C., Sanjuan-Hidalgo, J., García, D.J., Arenas, A.M., Matia-González, A.M., and Medina, P.P. (2023). SWI/SNF complexes in hematological malignancies: biological implications and therapeutic opportunities. Mol. Cancer *22*, 39. 10.1186/s12943-023-01736-8.

322. Wang, Z., Wang, P., Li, Y., Peng, H., Zhu, Y., Mohandas, N., and Liu, J. (2021). Interplay between cofactors and transcription factors in hematopoiesis and hematological malignancies. Signal Transduct. Target. Ther. *6*, 24. 10.1038/s41392-020-00422-1.

323. Liu, J., Zhang, J., Ginzburg, Y., Li, H., Xue, F., De Franceschi, L., Chasis, J.A., Mohandas, N., and An, X. (2013). Quantitative analysis of murine terminal erythroid differentiation in vivo: novel method to study normal and disordered erythropoiesis. Blood *121*, e43-49. 10.1182/blood-2012-09-456079.

324. Mei, Y., Liu, Y., and Ji, P. (2021). Understanding terminal erythropoiesis: An update on chromatin condensation, enucleation, and reticulocyte maturation. Blood Rev. *46*, 100740. 10.1016/j.blre.2020.100740.

325. Wall, L., deBoer, E., and Grosveld, F. (1988). The human beta-globin gene 3' enhancer contains multiple binding sites for an erythroid-specific protein. Genes Dev. *2*, 1089–1100. 10.1101/gad.2.9.1089.

326. Mak, K.S., Funnell, A.P.W., Pearson, R.C.M., and Crossley, M. (2011). PU.1 and Haematopoietic Cell Fate: Dosage Matters. Int. J. Cell Biol. *2011*, 808524. 10.1155/2011/808524.

327. Shaw, G.C., Cope, J.J., Li, L., Corson, K., Hersey, C., Ackermann, G.E., Gwynn, B., Lambert, A.J., Wingert, R.A., Traver, D., et al. (2006). Mitoferrin is essential for erythroid iron assimilation. Nature *440*, 96–100. 10.1038/nature04512.

328. Lamonica, J.M., Vakoc, C.R., and Blobel, G.A. (2006). Acetylation of GATA-1 is required for chromatin occupancy. Blood *108*, 3736–3738. 10.1182/blood-2006-07-032847.

329. Boyes, J., Byfield, P., Nakatani, Y., and Ogryzko, V. (1998). Regulation of activity of the transcription factor GATA-1 by acetylation. Nature *396*, 594–598. 10.1038/25166.

330. Wu, W., Cheng, Y., Keller, C.A., Ernst, J., Kumar, S.A., Mishra, T., Morrissey, C., Dorman, C.M., Chen, K.-B., Drautz, D., et al. (2011). Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. Genome Res. *21*, 1659–1671. 10.1101/gr.125088.111.

331. Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., et al. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. Publ. Protein Soc. *30*, 187–200. 10.1002/pro.3978.

332. Kumar, R.D., Searleman, A.C., Swamidass, S.J., Griffith, O.L., and Bose, R. (2015). Statistically identifying tumor suppressors and oncogenes from pan-cancer genome-sequencing data. Bioinforma. Oxf. Engl. *31*, 3561–3568. 10.1093/bioinformatics/btv430.

333. Lyu, J., Li, J.J., Su, J., Peng, F., Chen, Y.E., Ge, X., and Li, W. (2020). DORGE: Discovery of Oncogenes and tumoR suppressor genes using Genetic and Epigenetic features. Sci. Adv. *6*, eaba6784. 10.1126/sciadv.aba6784.

334. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. 10.1038/s41586-021-03819-2.

335. Tavanaei, A., Anandanadarajah, N., Maida, A., and Loganantharaj, R. (2017). A deep learning model for predicting tumor suppressor genes and oncogenes from PDB structure. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 613–617. 10.1109/BIBM.2017.8217722.

336. Grabher, C., von Boehmer, H., and Look, A.T. (2006). Notch 1 activation in the molecular pathogenesis of T-cell acute lymphoblastic leukaemia. Nat. Rev. Cancer *6*, 347–359. 10.1038/nrc1880.

337. Dotto, G.P. (2009). Crosstalk of Notch with p53 and p63 in cancer growth control. Nat. Rev. Cancer *9*, 587–595. 10.1038/nrc2675.

338. Villamar-Cruz, O., Loza-Mejía, M.A., Arias-Romero, L.E., and Camacho-Arroyo, I. (2021). Recent advances in PTP1B signaling in metabolism and cancer. Biosci. Rep. *41*, BSR20211994. 10.1042/BSR20211994.

339. Soussi, T., and Wiman, K.G. (2015). TP53: an oncogene in disguise. Cell Death Differ. *22*, 1239–1249. 10.1038/cdd.2015.53.

340. Yang, L., Han, Y., Suarez Saiz, F., and Minden, M.D. (2007). A tumor suppressor and oncogene: the WT1 story. Leukemia *21*, 868–876. 10.1038/sj.leu.2404624.

341. Jotte, M.R.M., and McNerney, M.E. (2022). The significance of CUX1 and chromosome 7 in myeloid malignancies. Curr. Opin. Hematol. *29*, 92. 10.1097/MOH.0000000000000699.

342. An, N., Khan, S., Imgruet, M.K., Jueng, L., Gurbuxani, S., and McNerney, M.E. (2023). Oncogenic RAS promotes leukemic transformation of CUX1-deficient cells. Oncogene *42*, 881–893. 10.1038/s41388-023-02612-x.

343. Maresca, M., van den Brand, T., Li, H., Teunissen, H., Davies, J., and de Wit, E. (2023). Pioneer activity distinguishes activating from non-activating SOX2 binding sites. EMBO J. *42*, e113150. 10.15252/embj.2022113150.

344. McBride, M.J., Pulice, J.L., Beird, H.C., Ingram, D.R., D'Avino, A.R., Shern, J.F., Charville, G.W., Hornick, J.L., Nakayama, R.T., Garcia-Rivera, E.M., et al. (2018). The SS18-SSX Fusion Oncoprotein Hijacks BAF Complex Targeting and Function to Drive Synovial Sarcoma. Cancer Cell *33*, 1128-1141.e7. 10.1016/j.ccell.2018.05.002.

345. Arnosti, D.N., and Kulkarni, M.M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J. Cell. Biochem. *94*, 890–898. 10.1002/jcb.20352.

346. Fang, B., and Roth, J.A. (2003). Tumor-suppressing gene therapy. Cancer Biol. Ther. *2*, S115-121.

347. Ueda, J., Yamazaki, T., and Funakoshi, H. (2023). Toward the Development of Epigenome Editing-Based Therapeutics: Potentials and Challenges. Int. J. Mol. Sci. *24*, 4778. 10.3390/ijms24054778.

348. Fine-tuning epigenome editors (2022). Nat. Biotechnol. *40*, 281. 10.1038/s41587-022-01270-w.

349. Sun, D., Luo, M., Jeong, M., Rodriguez, B., Xia, Z., Hannah, R., Wang, H., Le, T., Faull, K.F., Chen, R., et al. (2014). Epigenomic profiling of young and aged HSCs reveals concerted changes during aging that reinforce self-renewal. Cell Stem Cell *14*, 673–688. 10.1016/j.stem.2014.03.002.

350. Adelman, E.R., Huang, H.-T., Roisman, A., Olsson, A., Colaprico, A., Qin, T., Lindsley, R.C., Bejar, R., Salomonis, N., Grimes, H.L., et al. (2019). Aging Human Hematopoietic Stem Cells Manifest Profound Epigenetic Reprogramming of Enhancers That May Predispose to Leukemia. Cancer Discov. *9*, 1080–1101. 10.1158/2159-8290.CD-18-1474.

351. Ribeiro-Silva, C., Vermeulen, W., and Lans, H. (2019). SWI/SNF: Complex complexes in genome stability and cancer. DNA Repair *77*, 87–95. 10.1016/j.dnarep.2019.03.007.

352. Mouly, E., Chemin, K., Nguyen, H.V., Chopin, M., Mesnard, L., Leite-de-Moraes, M., Burlen-defranoux, O., Bandeira, A., and Bories, J.-C. (2010). The Ets-1 transcription factor controls the development and function of natural regulatory T cells. J. Exp. Med. *207*, 2113–2125. 10.1084/jem.20092153.

353. Yang, X.-P., Ghoreschi, K., Steward-Tharp, S.M., Rodriguez-Canales, J., Zhu, J., Grainger, J.R., Hirahara, K., Sun, H.-W., Wei, L., Vahedi, G., et al. (2011). Opposing regulation of the locus encoding IL-17 through direct, reciprocal actions of STAT3 and STAT5. Nat. Immunol. *12*, 247–254. 10.1038/ni.1995.

354. Rekhtman, N., Radparvar, F., Evans, T., and Skoultchi, A.I. (1999). Direct interaction of hematopoietic transcription factors PU.1 and GATA-1: functional antagonism in erythroid cells. Genes Dev. *13*, 1398–1411. 10.1101/gad.13.11.1398.

355. Hosokawa, H., Ungerbäck, J., Wang, X., Matsumoto, M., Nakayama, K.I., Cohen, S.M., Tanaka, T., and Rothenberg, E.V. (2018). Transcription Factor PU.1 Represses and Activates Gene Expression in Early T Cells by Redirecting Partner Transcription Factor Binding. Immunity *48*, 1119-1134.e7. 10.1016/j.immuni.2018.04.024.

356. Rodriguez, P., Bonte, E., Krijgsveld, J., Kolodziej, K.E., Guyot, B., Heck, A.J.R., Vyas, P., de Boer, E., Grosveld, F., and Strouboulis, J. (2005). GATA-1 forms distinct activating and repressive complexes in erythroid cells. EMBO J. *24*, 2354–2366. 10.1038/sj.emboj.7600702.

357. Guo, X., Plank-Bazinet, J., Krivega, I., Dale, R.K., and Dean, A. (2020). Embryonic erythropoiesis and hemoglobin switching require transcriptional repressor ETO2 to modulate chromatin organization. Nucleic Acids Res. *48*, 10226–10240. 10.1093/nar/gkaa736.

358. Griffin, C.T., Brennan, J., and Magnuson, T. (2008). The chromatin-remodeling enzyme BRG1 plays an essential role in primitive erythropoiesis and vascular development. Dev. Camb. Engl. *135*, 493–500. 10.1242/dev.010090.

359. Azad, P., Caldwell, A.B., Ramachandran, S., Spann, N.J., Akbari, A., Villafuerte, F.C., Bermudez, D., Zhao, H., Poulsen, O., Zhou, D., et al. (2022). ARID1B, a molecular suppressor of erythropoiesis, is essential for the prevention of Monge's disease. Exp. Mol. Med. *54*, 777–787. 10.1038/s12276-022-00769-1.

360. Letting, D.L., Rakowski, C., Weiss, M.J., and Blobel, G.A. (2003). Formation of a tissue-specific histone acetylation pattern by the hematopoietic transcription factor GATA-1. Mol. Cell. Biol. *23*, 1334–1340. 10.1128/MCB.23.4.1334-1340.2003.

361. Kim, Y.W., Kang, Y., Kang, J., and Kim, A. (2020). GATA-1-dependent histone H3K27 acetylation mediates erythroid cell-specific chromatin interaction between CTCF sites. FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. *34*, 14736–14749. 10.1096/fj.202001526R.