EPIGENETIC REGULATION IN NORMAL AND MALIGNANT HEMATOPOIESIS

Roger Mulet Lázaro

EPIGENETIC REGULATION IN NORMAL AND MALIGNANT HEMATOPOIESIS

ROGER MULET LÁZARO

ISBN: 978-94-6361-791-8
Layout: Egied Simons
Cover: Roger Mulet Lázaro
Printing: Optima Grafische Communicatie, Rotterdam

This thesis and all the supplementary materials are also available at:
https://hema13.erasmusmc.nl/thesisRML2022.html
The work described in this thesis was performed at the Department of Hematology at the Erasmus University Medical Center, Rotterdam, The Netherlands.
The printing of this thesis was financially supported by the Erasmus University Rotterdam and Oncode.

EPIGENETIC REGULATION IN NORMAL AND MALIGNANT HEMATOPOIESIS

EPIGENETISCHE REGULATIE IN NORMALE EN MALIGNE HEMATOPOËSE

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. A.L. Bredenoord

and in accordance with the decision of the Doctorate Board.
The public defence shall be held on

Thursday, 19 January, 2023 at 15.30 hrs

by

ROGER MULET LÁZARO

born in Barcelona, Spain

Erasmus University Rotterdam

DOCTORAL COMMITTEE

Promotor: Prof. dr. H.R. Delwel

Other members: Prof. dr. J. Cools
Prof. dr. I.P. Touw
Dr. K.S. Wendt

Co-promotor: Dr. B.J. Wouters

CONTENTS

CHAPTER 1

General introduction

1. INTRODUCTION

Introduction

"Science as a whole certainly cannot allow its judgment about facts to be distorted by ideas of what ought to be true, or what one may hope to be true." Conrad Waddington

Every day, the hematopoietic system produces billions of blood cells of multiple lineages, with essential functions in nutrient transport and immune defense. This process is strictly regulated by epigenetic factors, but disruptions in these mechanisms may lead to aberrant blood cell production. Acute myeloid leukemia (AML) is a malignant disorder characterized by the clonal proliferation of immature myeloid precursors in the bone marrow, at the expense of the production of healthy blood cells [1]. Rather than as a single entity, AML is often defined as a heterogeneous group of diseases, where are identified by particular genetic or cytogenetic aberrations. In the last years, it has become increasingly clear that epigenetic dysregulation is a major contributor to the pathogenesis of AML: most patients carry mutations in epigenetic modulators [2], and epigenetic marks such as methylation can also identify biologically distinct subtypes [3]. Furthermore, changes (genetic or epigenetic) in regulatory elements have been identified as drivers of altered gene expression [4].

The present thesis attempts to further the understanding of epigenetic control of both healthy and malignant hematopoiesis. Therefore, this introductory section covers fundamental concepts in the fields of hematopoiesis, epigenetics and leukemogenesis, organized in the following subsections:

- The hematopoietic system
- Epigenetic control of hematopoiesis
- Acute myeloid leukemia
- Scope and aims of this thesis

2. THE HEMATOPOIETIC SYSTEM

The hematopoietic system encompasses a wide range of cell types, which are classically classified in two major lineages: the myeloid branch, responsible for nutrient transport, hemostasis and innate immunity, and the lymphoid branch, mainly involved in adaptive immunity [5]. In adult humans, all hematopoietic cells are born in the bone marrow, but those in the lymphoid lineage migrate to lymphatic organs, such as the spleen and the thymus, to complete their maturation. On the contrary, myeloid cells (the word myeloid derives from the Greek *muelós*, which means "marrow") arise from the bone marrow as fully differentiated cells.

In the myeloid lineage, erythrocytes transport oxygen throughout the body, whereas megakaryocytes give rise to platelets. Granulocytes (comprising neutrophils, basophils and eosinophils), monocytes (which become macrophages upon migration to tissues) and dendritic cells are myeloid cells that participate in innate immunity and contribute to the initiation of adaptive responses. The lymphoid branch contains B and T lymphocytes, as well as natural killer (NK) cells, which carry out adaptive immune responses. Recent advances, however, have blurred the boundaries between the roles of these two lineages, as shown by the existence of innate lymphoid cells [6].

Because mature blood cells are predominantly short lived, the maintenance of these functions requires the daily production of more than 500 billion blood cells every day in adult humans [7]. This remarkable feat is made possible by the process known as **hematopoiesis**, in which hematopoietic stem cells (HSCs) replicate and specialize into progenitor cells and finally into functioning mature phenotypes [8].

2.1 The hematopoietic stem cell

HSCs are the only bone marrow cells **capable of differentiating into all blood cell lineages (multipotency) and replicating into other HSCs (self-renewal)** [9]. These two central properties were originally proposed by James Till and Ernest McCulloch in the early 1960s [10]. In the course of their research on radiation sensitivity of normal mouse bone marrow, they identified a class of cells "capable of continued proliferation [and] differentiation" that they originally named "spleen colony-forming units" (CFU-S) [11]. Later on, they conclusively demonstrated the clonal origin of these cells by introducing and tracking unique chromosomal aberrations prior to transplantation into other mice [12]. In a subsequent study, they showed that these CFU-S possessed the ability of self-renewal, which led the authors to conclude they were bona fide HSCs [13].

Experimentally, **HSCs are defined by their ability to reconstitute the entire blood system**, which is typically shown by engrafting in lethally irradiated recipients and establishing long-term multi-lineage hematopoiesis [9]. In fact, it has been proven that a single cell can give rise to a whole hematopoietic system [14]. However, it is not always convenient or feasible to conduct an *in vivo* transplantation assay to assess multipotency, so currently HSCs are

prospectively isolated and purified by fluorescence-activated cell sorting (FACS) on the basis of known HSC surface markers [8]. Spangrude and colleagues were the first to purify a bone marrow-reconstituting population of HSCs using surface marker staining, characterized by Thy-1^{lo} Lin^- $Sca\text{-}1^+$ in mouse [15].

Successive experiments showed that the mutipotent compartment of the bone marrow comprises, in fact, multiple populations. In 1994, Morrison and Weissman characterized Long-Term (LT)-HSC, Short-Term (ST)-HSC and Multi-Potent Progenitors (MPP), each endowed with different degrees of self-renewal and multipotency, as well as different surface markers [16]. Originally described in mice, these populations were eventually found in humans too, but with important differences in their surface markers. One major particularity of human HSCs is that they exhibit CD34 surface expression, in contrast with $CD34^{-/low}$ LT-HSC in mice [17]. $CD34^+$ bone marrow cells constitute 1.5% of all mononuclear cells in the marrow, and 0.05% in peripheral blood [18]. Aside from its role as a surface marker, CD34 participates in the migration of HSC and other progenitor cells (HSPCs) via interaction with vascular selectins [19]. However, a rare HSC population devoid of CD34 expression ($CD34^-$) also exists in humans, with immature and quiescent characteristics [20]. Since these $C34^-$ cells are able to self-renew and differentiate into $CD34^+$ LT-HSCs [21], it has been speculated they could represent a last reservoir of HSCs for situations of stress [22].

In the last two decades, several additional markers have been identified in both human and mouse to more accurately define subsets of HSCs and multipotent progenitors [23]. In humans but not in mice, the combination $CD34^+CD38^-$ provides further enrichment for HSCs compared to $CD34^+$ alone [24,25]. The presence of CD90 (Thy-1) is also associated with HSCs [26], whereas CD45RA [27] delineates more differentiated progenitors. The signaling lymphocyte activation molecule (SLAM) family of receptors —including CD150, CD244, and CD48— can be used for isolation of HSCs in mice, but lack discriminatory power in humans [24].

2.2 The hematopoietic hierarchy: differentiation and self-renewal

A critical observation by Till and McCulloch was the need for a balance between self-renewal and differentiation in HSCs. Disruptions in this balance may lead to adverse outcomes, including depletion of the HSC compartment due to insufficient self-renewal, or leukemia if differentiation is blocked. In 1964 they published a stochastic model describing the two possibilities that every HSC faces upon division (either differentiation or self-renewal) in analogy with the decay of radioactive nuclei [28]. To this day, the dilemma between self-renewal and differentiation remains a focus of intensive research: how can cells retain sufficient HSCs while at the same time meeting the enormous demand for new mature blood cells? Even more strikingly, the frequency of HSCs in the bone marrow is estimated around 0.01% in mice, even less in humans [29], and yet billions of blood cells are produced on a daily basis.

The answer to this conundrum lies in the multi-tiered nature of hematopoiesis, often described as a **hierarchical structure in which multipotency is progressively restricted** [30].

According to this paradigm, hematopoietic differentiation is a tree-like branched roadmap that occurs in a stepwise manner, with HSCs at the apex (Figure 1). As outlined above, HSCs initially give rise to MPPs that retain full-lineage differentiation potential, but lose their self-renewal ability. Further downstream, MPPs advance to oligopotent progenitors: common lymphoid progenitors (CLPs) and common myeloid progenitors (CMPs). Subsequently, these oligopotent progenitors give rise to lineage-committed effector cells, which in turn may differentiate further into fully specialized cells. For example, CMPs give rise to megakaryocyte/erythrocyte progenitors (MEPs) and granulocyte/macrophage progenitors (GMPs) [30]. The transitions between states are driven by changes in key transcription factors (TFs) that are involved either in HSC maintenance or in lineage specification [8].

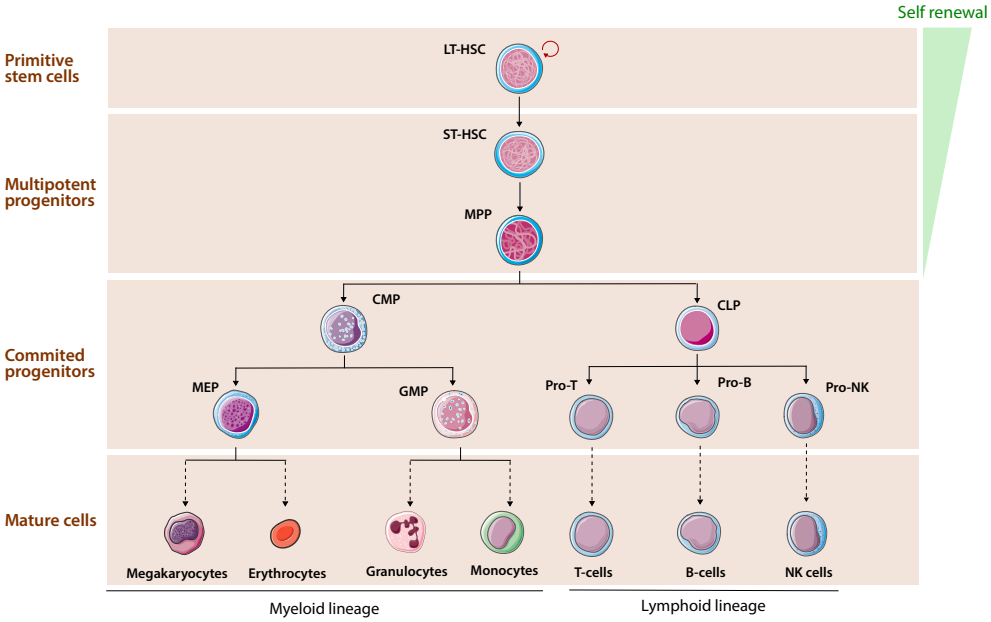


Figure 1. Traditional depiction of hematopoietic differentiaton as a hierarchical tree. This diagram has been adapted from [8,17] using material from Servier Medical Art by Servier, licensed under a Creative Commons Attribution 3.0 unported license.

This multi-tiered structure allows LT-HSCs to remain quiescent to minimize exhaustion and cell cycle-associated DNA damage, relying instead on ST-HSCs and downstream progenitors to carry most of the replicative burden necessary for steady-state hematopoiesis [5]. Approximately 75% of LT-HSC are estimated to be in the G0 phase of the cell cycle at any time, with 5% in the S/G2/M phases and 20% in G1 [31]. This quiescent state is regulated by a number of intrinsic and extrinsic signals. Among the former, the transcription factors PU.1 (*SPI1*) and SATB1 limit proliferation by modulating the expression of multiple cell cycle regulators [32]. Besides, HSCs receive extrinsic cues from cells in their microenvironment (the so-called "niche", see section 2.4) that enforce quiescence. For example, it has been shown that blockade of TGFβ signaling delays the return of cycling HSCs to quiescence *in vivo* [33].

In situations of stress, such as a serious infection or blood loss, HSCs can become activated to proliferate and differentiate [34]. After the damage is repaired, activated HSCs return to dormancy, indicating that the switch between the two states is not purely stochastic, but a well regulated physiological response. Repeated or continuous exit from dormancy has been linked to DNA damage and attrition, which provides an explanation for accumulation of DNA damage with ageing [35]. A correct understanding of this process is critical not only because of its possible implication in ageing or disease, but also for its relevance in therapy. Granulocyte colony-stimulating factor (G-CSF) is used clinically to treat leukopenia and for peripheral blood mobilization of HSPCs, which is currently the preferred option for transplantation [36]. Initially, G-CSF was reported to promote the proliferation of LT-HSC cells, some of which would then migrate to the periphery [37]. Increased exhaustion derived from this process was a possible concern associated with this practice, but recent evidence indicates that G-CSF mobilizes dormant HSCs without proliferation [38]. The downside could be, however, that the bone marrow of the donor becomes depleted of LT-HSC.

Although all HSCs are defined by their ability to self-renew and differentiate, an increasing body of evidence shows that these properties are not equally shared by all clones. Early studies with retrovirally marked HSCs had identified lineage-restricted repopulation patterns [39], but the idea of distinct stem cell classes was rejected in favor of dynamic changes of lineage contribution after transplantation [40]. A long-standing view was that HSCs were homogeneous, but individual clones could behave differently depending either on environmental cues or stochastic mechanisms [41,42]. However, Sieburg and colleagues showed that the HSC compartment consists of distinct subsets with varying potentials for self-renewal and differentiation, which are epigenetically programmed [43,44]. Dykstra et al. defined four subsets (α, β, γ and δ) with different degrees of self-renewal and preference for myeloid or lymphoid lineages [45]. More recently, the Goodell group identified stable myeloid- (My) and lymphoid-biased (Ly) HSCs that are differentially regulated by TGFβ1 [46].

2.3 Evolving views of hematopoiesis

Traditionally, hematopoietic differentiation has been portrayed as **a series of stepwise transitions between discrete cell states** [8]. At each stage of differentiation, populations are functionally homogeneous and only commit further by undergoing binary fate decisions (Figure 2A). However, this "classical" model has come under scrutiny in the last decade as more evidence accumulates that the hematopoietic hierarchy consists of heterogeneous populations with gradual progression from one to the next [23]. The insights derived from new technologies, particularly single cell sequencing, have been critical in this paradigm shift.

The use of surface markers and flow cytometry enabled the distinction between different stages of maturation that underpins the classical model of hematopoiesis. While this strategy shed light on this complex process, the limited availability of surface markers initially led to oversimplification. As new markers were discovered and techniques like mass cytometry

were developed, models grew in complexity and the notion of a "hematopoietic continuum" began to emerge [47]. Nevertheless, the usefulness of flow cytometry is dependent on the existence of specific surface markers in a given cell type. In contrast, massively parallel single cell RNA-sequencing (scRNA-seq) allows unbiased characterization of thousands of cells and their representation by similarity of gene expression [48,49]. Pioneering scRNA-seq studies in mouse [50,51] and humans [52–54] showed a continuous distribution of cells ranging from primitive HSCs to terminally differentiated stages.

Taken together, these findings have converged on a revised model of the **hematopoietic tree in which lineage commitment is a continuous process** (Figure 2B) [23]. In this new framework, the progenitors of the classical model reflect arbitrary groups of cells delimited by certain surface markers rather than true stable cell states. This proposed continuum accommodates the heterogeneity of the HSC compartment described above, as well as the multiple MPP subpopulations identified by others [55]. Nonetheless, there are caveats with this model. On the one hand, despite their limitations, surface markers identify subpopulations with unique functions and restricted differentiation potential [23]. On the other, coupling scRNA-seq with genetic perturbations identified distinct transition points safeguarded by key hematopoietic TFs [56,57]. Furthermore, variability of gene expression may cause enough transcriptome diversity to blur the differences between phenotypically similar cells [58]. To reconcile these different lines of evidence, **a compromise has emerged in the shape of a "punctuated continuum"** where punctuated transitions delineate groups of functionally similar cells in the differentiation continuum (Figure 2C) [5,23].

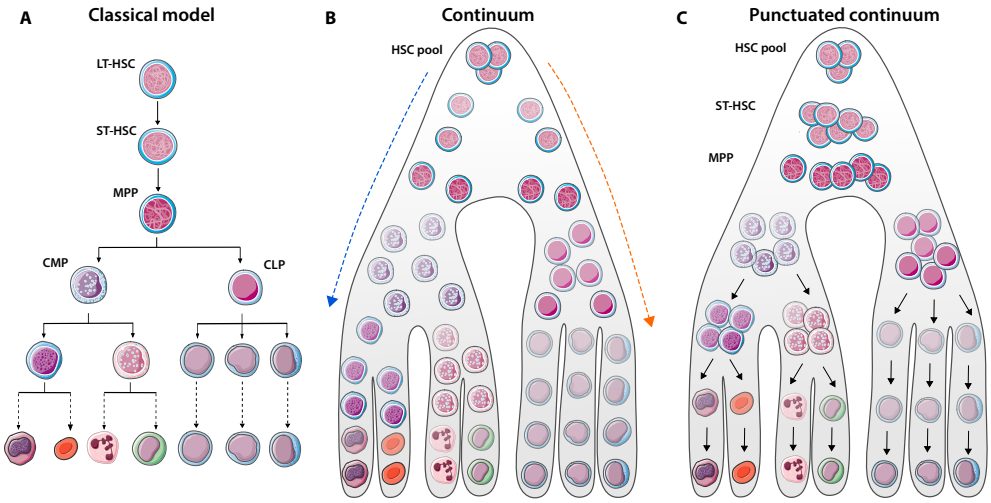


Figure 2. Evolving view of hematopoiesis. In the classical model of hematopoiesis, differentiation progresses in a step-wise manner (A), in contrast with the hematopoietic continuum inferred from single cell data (B). The punctuated continuum model reconciles the two views, adding transition stages between groups of functionally similar cells (C). Figure adapted from [5] and [23] using material from Servier Medical Art by Servier.

Another major development in the last decade is the ability to study hematopoiesis under physiological conditions (steady state), thanks to lineage-tracing approaches [59]. Previously, most evidence derived from *in vitro* colony assays and *in vivo* transplantation experiments, which fail to reproduce the environment and stimuli found in unperturbed bone marrow. A landmark study using the Sleeping Beauty transposase to tag progenitor cells and their progeny revealed **that LT-HSCs have a limited contribution to blood production during adulthood** [60]. Instead, the authors proposed that a large pool of ST-HSCs and MPPs specified in the perinatal period support hematopoiesis, leading to substantial clonal diversity. Nevertheless, another study with YFP-based label tracing concluded that, while most mature cells are indeed derived from ST-HSCs, a continuous input from LT-HSCs is required [61]. The Reizis group showed a much higher and faster contribution of LT-HSCs to steady state hematopoiesis, with 3-8% of these cells differentiating every day [62].

Furthermore, a combination of cell barcoding and scRNA-seq revealed a number of surprising findings: HSCs and early progenitors exhibit lineage biases, and multiple routes converge on monocyte differentiation [63]. Crucially, other studies also showed the predominance of unipotent cells within compartments that are multipotent as a whole [53]. At the same time, these compartments retain bilineage potential until late stages of hematopoiesis, calling into question the traditional divide between myeloid and lymphoid branches [64].

In summary, new technologies have reshaped the traditional conception of hematopoiesis, and will probably continue to do so in the years to come. Even though many unknowns remain, it has become clear that hematopoiesis is not a series of discrete and compartmentalized stages, and that fate choices are not purely binary decisions.

2.4 The influence of the niche

Hematopoietic stem cells do not reside in the marrow in isolation. As early as 1978, drawing from findings by Michael Dexter [65] and his own observations, Ray Schofield proposed that HSCs are surrounded by other cells that prevent their maturation and ensure their proliferation, forming a microenvironment that he termed “the niche” [66]. In this novel framework, differences between the bone marrow niche and the spleen could explain the diminished self-renewal of CFU-S cells compared to bone marrow HSCs. Since then, advances labelling and imaging technologies have made it possible to characterize in detail this microenvironment [67].

Multiple cell types residing in the niche form a complex multicellular network essential for HSC localization, maintenance and differentiation, including osteoblasts[68], CXCL12-abundant reticular cells (CAR cells), macrophages, megakaryocytes [69] and mesenchymal stem cells (MSCs) [70]. Several factors modulate HSC function, but **SCF/KITLG, CXCL12 and thrombopoietin** are known to be absolutely critical for their survival [71]. CXCL12 controls the bone marrow retention of HSCs through interaction with the CXCR4 expressed on the

surface, but it can also mediate their mobilization [72]. The main sources of SCF and CXCL12 in the bone marrow are **perivascular and endothelial cells** located around sinusoids, most of which express the leptin receptor [73]. These cells, which have also been described as "CAR cells", represent 0.3% of the bone marrow cells and are critical for HSC maintenance, according to ablation experiments [71].

Alterations in this multicellular network inevitably lead to disturbances in normal HSC function**.** During ageing, changes in the bone marrow microenvironment further contribute to the intrinsic deterioration of HSCs [74], as shown by the lower transplantation efficiency of HSCs in aged recipients [75]. Aged perivascular epithelial cells exhibit increased leakiness and express lower levels of SCF and CXCL12, a process that can be reversed by infusion of young cells [76]. On the other hand, an altered bone marrow niche can promote and sustain malignant transformation [70]. Deletion of *Dicer1* in mouse bone progenitors fostered the development of myelodysplastic syndrome (MDS), which occasionally transformed to AML [77]. Conversely, leukemic stem cells (LSCs) can remodel their microenvironment to better suit their needs. In myeloproliferative neoplasia, MSCs overproduce altered osteoblast progenitors, which exhibit reduced ability to sustain normal HSCs, but support LSCs [78]. In a model of AML with MLL-AF9, degeneration of sympathetic nerve fibres also lead to increased osteoblast differentiation at the expense of HSC-supporting cells [79].

3. EPIGENETIC CONTROL OF HEMATOPOIESIS

Hematopoiesis is a tightly regulated process that ensures a steady supply of blood cells of multiple lineages, each of them endowed with specialized functions. This process is punctuated by a series of choices between cell fates, forming a continuous spectrum of progressively differentiated cell types. The choices along this continuum are governed by intrinsic and extrinsic factors, which are collectively known as "epigenetics". The prefix "epi" in Greek means "on" or "above", and thus the term "epigenetics" refers to factors in the genome beyond the genetic code. It was originally coined by the embryologist Conrad Waddington in 1942 to define the "whole complex of developmental processes" that bridge "genotype and phenotype" [80]. He went on to define the "**epigenetic landscape**" as a number of developmental pathways that a cell may take during differentiation, metaphorically depicted as a ball rolling downhill through a series of branching ridges and valleys [81]. Each valley represents a possible cell fate and the ridges between the valleys are barriers that maintain that fate once it has been chosen, often in a binary manner [82]. The shape of the landscape is determined by genetic regulatory mechanisms, which are illustrated as pegs underpinning the ridges and valleys [83]. These visual metaphors, more than 50 years old, are reminiscent of our current understanding of hematopoietic differentiation.

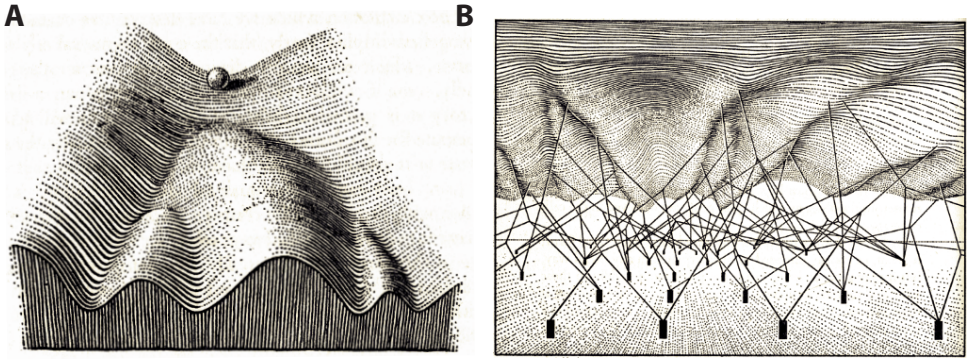


Figure 3. Early representations of the epigenetic landscape. (A) A cell, represented as a ball, rolls down the landscape through a series of branching points representing fate choices. (B) Underlying regulation by genes, represented as pegs underpinning the landscape [81]

Although the epigenetic landscape proposed by Waddington –only few years after the discovery of the double helix structure of DNA– was strikingly prescient, the processes bridging genotype and phenotype were merely an abstraction. The first instance of gene regulation can be traced back to the "controlling elements" proposed by Barbara McClintock, based on her discoveries of transposition in maize [84]. In 1961, the operon model of Jacob and Monod firmly established the existence of gene regulation [85]. Only decades later would advances in molecular biology reveal mechanisms such as DNA methylation, histone modifications or chromatin conformation – initially discovered and researched independently. Over the

years, “epigenetics” has become an umbrella term for any mechanism governing gene expression without changes in the DNA sequence [86].

Increased insight into these mechanisms has been critical in the characterization of hematopoiesis, both in health and in disease. We now know that transitions between differentiation stages are a result of changes in gene expression following epigenetic mechanisms, which are thus ultimately responsible for fate choices [87]. The current section describes these mechanisms, starting with general principles of gene expression and continuing with their involvement in hematopoiesis.

3.1 Principles of transcriptional regulation

The completion of the Human Genome Project in 2004 revealed that the human genome encodes for roughly 20,000 protein-coding genes [88]. Although many of these genes are ubiquitously expressed, some are specific to a certain tissue or even certain cells [58,89]. These tissue-specific genes are responsible for morphological and functional differences between cells in development and differentiation, and as such they are under strict regulation.

Gene expression starts with transcription, defined as the copying of a DNA sequence into RNA by a member of the RNA polymerase family of enzymes. RNA polymerase II (Pol II) transcribes all protein-coding and most non-coding genes, whereas Pol I and Pol III transcribe ribosomal RNA (rRNA) and certain small non-coding RNAs (ncRNAs), respectively [90]. Transcription can be divided into three distinct phases: initiation, elongation and termination. Transcription begins at the **transcription start site** (TSS), located at the 5’ end of a gene, and progresses towards its 3’ end. The region around the TSS is known as the **promoter,** which comprises a proximal region upstream of the TSS and a core promoter of 40-50 base pairs (bp) around the TSS [91]. The proximal promoter region contains binding sites for **transcription factors (TFs)**, proteins that recognize specific DNA sequences and foster recruitment of the transcription machinery, including Pol II and other cofactors [91]. The **core promoter** serves as a platform for Pol II and the general transcription factors (GTFs), which together form the **pre-initiation complex** (PIC) [92].

Core promoters are generally sufficient to initiate transcription, but they have low basal activity. This activity can be further increased by interaction with another class of distal regulatory elements termed **enhancers** (Figure 4A) [92]. Enhancers are small segments of DNA that recruit TFs through short, specific DNA sequences (motifs) to regulate transcription [93]. They collaborate in the recruitment of Pol II by forming loops with target promoters, which explain their ability to act at a distance [94]. In addition, there are a number of other distal *cis*-regulatory elements (CREs) that participate in gene regulation, including silencers and insulators. Silencers reduce transcription from their target promoters by bringing repressive TFs, known as repressors [95]. Insulators bind architectural proteins that generate loop domains, thus blocking interaction across domains and favouring those within the same loop (Figure 4B) [96].

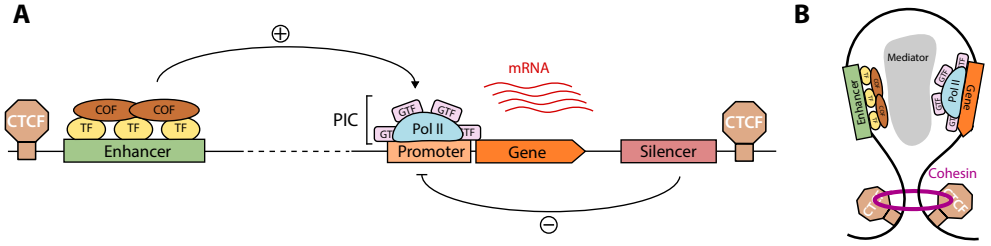


Figure 4. Transcription initiation is regulated by enhancers and promoters. (A) Promoters recruit general transcription factors (GTFs), which in turn facilitate the binding of RNA polymerase II (Pol II), leading to the formation of the pre-initation complex (PIC). Transcription from promoters is favoured by distal enhancers, which bind sequence-specific transcription factors (TF) and cofactors (COF). Figure adapted from [92]. (B) Chromatin loops enable contacts between distant enhancers and promoters, while preventing interactions with CREs outside the loop.

Moreover, the activation of gene regions is regulated by the three-dimensional organization of DNA. In eukaryotes, genomic DNA is hierarchically packaged by histones into **chromatin,** composed of strings of nucleosomes, to fit inside the nucleus and control its accessibility [97]. Chromatin can switch between transcriptionally active **euchromatin** and inactive **heterochromatin** [98]. Transcription requires binding of TFs and the transcriptional machinery to DNA, which is only possible if chromatin is open. Therefore, a number of factors such as DNA methylation and histone modifications regulate DNA accessibility by modulating the properties of nucleosomes, thus allowing transcription [97]. Proteins known as "chromatin remodellers" directly influence chromatin structure when recruited by pioneer TFs, which can uniquely bind the DNA in nucleosomes at enhancers and promoters [99].

3.2 Hierarchical folding of chromatin

1.1.1 The nucleosome

The basic unit of chromatin organization is the **nucleosome** (Figure 5). Nucleosomes were originally described by Don and Ada Olins in 1974 as a series of "beads on a string" observed by electron microscopy [100], and their structure was subsequently elucidated by Kornberg and Thomas in the same year [101]. Currently, it is well established that nucleosomes comprise i) a core particle of 146-147 base pairs of DNA wrapped in 1.65 superhelical turns around an octamer of histone proteins (two each of H2A, H2B, H3 and H4), ii) linker DNA of variable length between 10 to 90 bp [102]. Histones are small, positively charged proteins that can strongly bind the negatively charged backbone phosphates of DNA through electrostatic interactions [103]. This fiber of approximately 11 nm of diameter constitutes the primary level of organization of chromatin. It is estimated that 75–90% of genomic DNA is wrapped in nucleosomes [104].

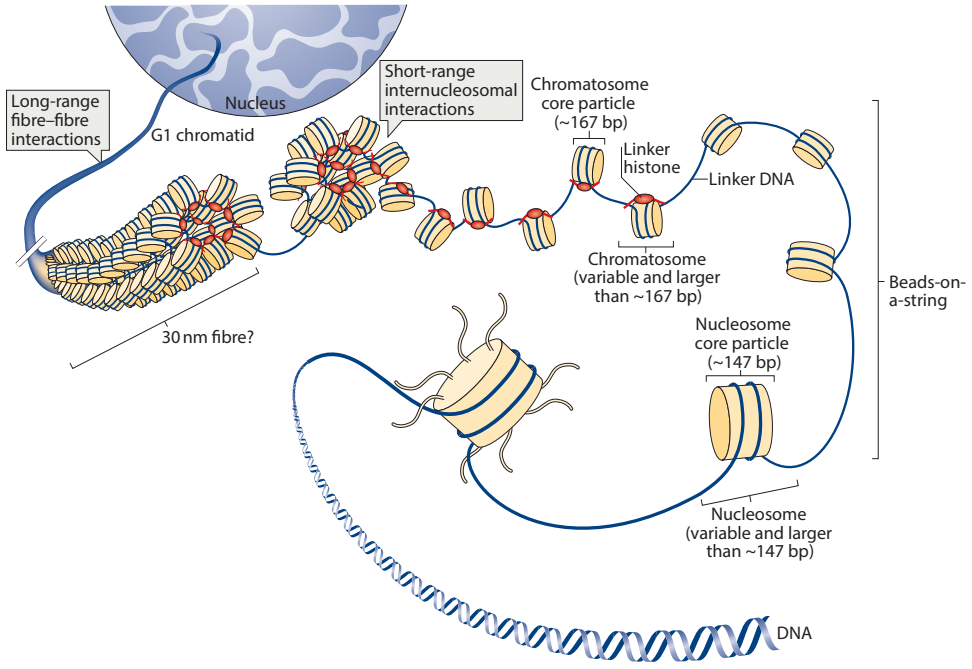


Figure 5. Schematic representation of hierarchical folding of chromatin, with the nucleosome as its basic unit [103]. Note that the 30 nm fiber has only been observed *in vitro* and its existence is therefore questioned.

However, nucleosomes only lead to a DNA compaction of 5-fold, whereas encapsulating the roughly 2 m of DNA in the nucleus of human cells requires a compaction of 10,000-fold [105]. This strongly hints at the presence of higher orders of chromatin organization. Indeed, binding of a linker histone H1 to 10 bp of DNA linker on both sides of the nucleosome results in further packaging of DNA in the form of the "chromatosome" [103]. For decades, it was widely accepted that a second level of organization was the a fiber of 30 nm *observed in vitro*, driven by nucleosome-nucleosome interactions [106]. Nevertheless, such a regular structure has not been observed *in vivo* [107], suggesting that an extensive 30-nm fiber is not stable in physiological conditions [108]. Instead, super-resolution microscopy revealed that nucleosomes form clusters or "clutches" of varying size [109], which indicates that chromatin may exhibit different folding levels at a smaller scale.

The function of DNA packaging in nucleosomes is twofold – in addition to ensuring that the DNA fits inside the nucleus, it plays a critical role in the regulation of gene expression [105]. Nucleosomes hinder the binding of sequence-specific TFs to CREs, including promoters and enhancers [110]. Furthermore, the wrapping of DNA around histones prevents transcriptional initiation by blocking the formation of the PIC at the TSS, while allowing chain elongation [111]. DNA wrapping around nucleosomes achieves efficient repression of coding genes and, perhaps more importantly, of the multitude of intergenic TSSs spread throughout the genome, thus limiting pervasive transcription [105]. Elongation can overcome nucleosome

barriers by displacing H2A/H2B dimers, while an hexasome remains attached, a process that can be aided by the FACT complex [112]. As transcription rate increases, higher density of Pol II may lead to complete eviction of the histone hexamer [113].

3.2.2 Nucleosome free regions

Nucleosome eviction or destabilization in **nucleosome-free regions (NFR)** is a critical requirement for the binding of TFs to *cis* regulatory elements and initiation of transcription [114]. These accessible chromatin regions are susceptible to digestion by nucleases, and as such they are also known as DNase hypersensitive sites (DHS) [110]. Galas and Schmitz developed **DNAse footprinting** as a method to study the fragments of DNA protected from DNase degradation by bound proteins – the "footprint" of such proteins [115]. Use of these methods revealed that active chromatin coincides with nuclease hypersensitivity, which is lost when loci return to an inactive state [116]. Moreover, the observation that the 5' ends of heat shock genes are hypersensitive to DNase I also suggested that sites in those regions may be recognized by regulators of gene expression [117]. Next generation sequencing has enabled DNA footprinting at the genome scale with single nucleotide resolution [118], a technique known as DNAse-seq.

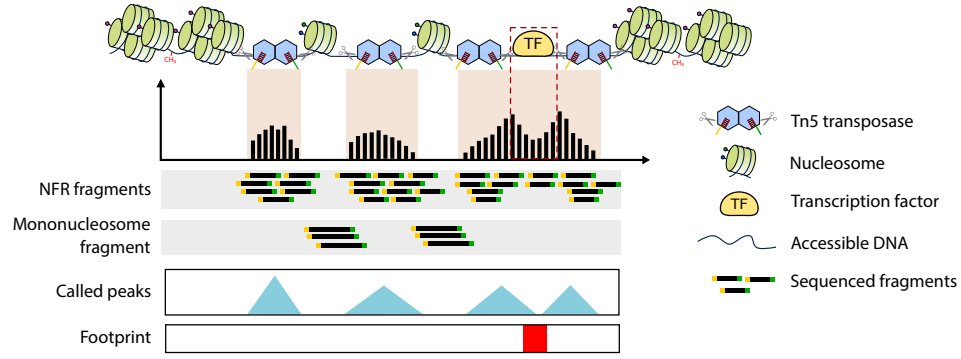


Figure 6. Assessing DNA accessibility with ATAC-seq (adapted from [125]). The Tn5 transposase cuts open chromatin and ligates adapters, enabling the sequencing of NFR fragments, which can be identified by peak calling algorithms.

DNAseq-seq identifies accessible nucleosome-depleted regions as peaks enriched for sequencing reads, in which narrow depressions correspond to TF footprints protected from DNAse [119]. More recently, the Assay for Transposase Accessible Chromatin with high-throughput sequencing (ATAC-seq) has gained prominence as an alternative to DNAse-seq. This technology probes DNA accessibility with hyperactive Tn5 transposase, which inserts sequencing adapters into open chromatin regions (Figure 6) [120]. Whereas DNAse-seq requires tens of millions of cells as input material, ATAC-seq can be performed on a few thousand cells, allowing the characterization of rare samples or cell populations. This, together with the fact that new ATAC-seq protocols are cheaper, less laborious and deliver similar quality as DNAse-seq, has led to a widespread adoption of the technique [121]. For example, genome-wide profiling of the chromatin accessibility landscape was conducted with ATAC-seq in 23

cancer types [122]. Although ATAC-seq and DNAseq-seq have different biases footprint shapes, they identify similar TF binding sites [123]. A variation of this technique, single-cell ATAC-seq (scATAC-seq), provides insights into cell-to-cell variation of the regulatory landscape [124].

3.2.3 Mechanisms of nucleosome eviction

Chromatin accessibility is facilitated by several processes, including the replacement of canonical histones with histone variants, the eviction or repositioning of histones by chromatin remodelers and the covalent modification of histones [126]. Aside from these active mechanisms, the underlying DNA sequence plays a critical importance in the determination of nucleosome positioning [127].

Although positively charged histones can interact with negatively charged DNA regardless of its base composition, **some sequences (such as TATA and CAG) have a much higher affinity for nucleosomes** than others [128]. On the contrary, homopolymeric tracts of poly(dA)-poly(dT) are intrinsically rigid and as such disfavor nucleosome formation [129], which is why they are often found in linker DNA between nucleosome core particles [127]. Moreover, these tracts are also present in gene promoters, where they stimulate transcription [130]. Although these fundamental principles were originally discovered in yeast, more recent studies in humans have also established a clear relationship between sequence and nucleosome positioning in human cells [131,132].

However, DNA sequence is not the major determinant of nucleosome positioning. Comparative studies of nucleosome assembly on genomic DNA show that the level of depletion in promoters is smaller *in vitro* than *in vivo*, indicating that other factors contribute to this depletion [133,134]. In particular, **ATP-dependent chromatin remodeling enzymes** are critical for the establishment of *in vivo* patterns of nucleosome positioning [127]. These proteins, commonly referred to as "remodelers", are classified into four functionally related subfamilies: imitation switch (ISWI), chromodomain helicase DNA-binding (CHD), switch/sucrose non-fermentable (SWI/SNF) and INO80 [135]. All of them share an ATP-dependent DNA translocase domain that binds the nucleosome and breaks contacts with the DNA, thereby promoting DNA translocation and nucleosome repositioning or editing [135]. In addition, they harbor additional domains that tailor the DNA translocation to specific functions and determine their selectivity for certain genomic locations.

The recruitment of remodelers is primarily mediated by transcription factors in a sequence-specific manner [93]. **Pioneer factors**, a special class of TF that can bind closed chromatin, collaborate with remodeling factors to make chromatin accessible to other TFs [136]. Cirillo and colleagues first used this term to describe the factors FOXA (HNF3) and GATA4, after demonstrating they bind nucleosome arrays and open compacted chromatin [137]. Previously, it had been shown that GATA4 is among the first TFs to bind essential regulatory sites in development [138] and that GATA1, a related member of the same family, induces the disruption of nucleosomes and formation of DHS [139]. The effect of GATA1, a transcriptional activator of erythroid-specific enhancers, is mediated by the recruitment of the BRG1 ATPase subunit of the SWI/SNF complex [140].

Moreover, remodelers can also be recruited and modulated by histone modifications via specific domains [135,141]. For example, acetylated lysines can be recognized by bromodomains present in subunits of the SWI/SNF complex, thus anchoring SWI/SNF to acetylated nucleosomes [142,143], which are displaced as a result [144]. This explains the ordered recruitment of these factors observed in promoters [145]. Conversely, other modifications shield nucleosomes against remodelers. For instance, the Polycomb repressive complex 1 (PRC1) inhibits chromatin remodeling by excluding SWI/SNF [146].

3.2.4 Higher order chromatin organization

Beyond the level of individual nucleosomes, chromatin is further organized into higher-order structures that play a critical role in gene regulation (Figure 7). Together with microscopy-based approaches, the chromosome conformation capture (3C) technology [147] and its derivatives (Box 1) have greatly improved our understanding of this hierarchical organization [148].

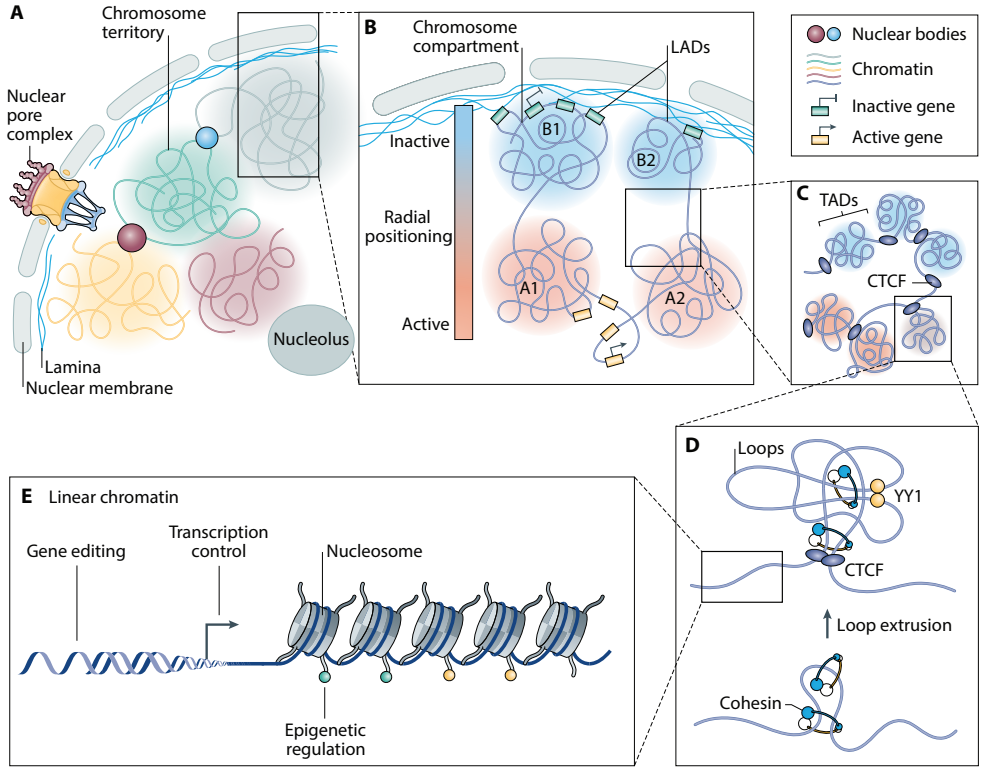


Figure 7. Three-dimensional genome organization and gene regulation [149]. Chromosomes are spatially segregated into subnuclear territories **(A)**, each of which contains two types of chromatin referred as "A" (active) or "B" (inactive) compartments **(B)**. At a scale of 10-100 kb, chromatin folds in topologically associating domains (TADs) delimited by CTCF **(C)**, which in turn harbor loops between enhancers and promoters **(D)**.

At the largest scale, **chromosomes segregate into independent "territories**" (Figure 7A), which were originally proposed more than a century ago, but whose existence was only validated by FISH in the 1980s [150]. The development of Hi-C revealed that chromosomes are folded into two **largely independent compartments** (Figure 7B), arbitrarily labeled A and B, defined by the eigenvector or first component of a principal component analysis (PCA) of the interaction matrix [151]. Compartment A, which roughly corresponds to euchromatin, is associated with greater gene density, active transcription and open chromatin, as well as active histone modifications. On the contrary, compartment B is more densely packed and exhibits a transcriptionally repressed state characteristic of heterochromatin. Regions belonging to one compartment preferentially interact with other regions of the same compartment, which are not necessarily contiguous in the linear genome. Spatial hubs bring together regions from different chromosomes, concentrating around nuclear speckles if they are transcriptionally active, or close to the nucleolus if they are inactive [152]. More recently, higher resolution Hi-C has suggested that compartments A and B can be subdivided into two and four subcompartments, respectively [153].

Box 1. Proximity ligation approaches to study chromatin architecture

Proximity ligation was first used to study DNA loops between the rat prolactin promoter and a distal enhancer in 1993 [154], but the key breakthrough introduced by the **3C technique** was the addition of formaldehyde crosslinking to improve the efficiency of proximity ligation reactions [147]. As a result, 3C enabled the detection of long-range chromatin interaction between any pair of genomic loci (one versus one). The 3C protocol starts with the formaldehyde treatment to crosslink the chromatin proteins to their associated DNA, followed by restriction enzyme digestion. The digested DNA is re-ligated in conditions that favor ligation of adjacent DNA, which is quantified by either PCR or sequencing approaches to determine chromatin contacts.

There are multiple variations of this protocol. In chromosome conformation capture-on-chip (**4C**), a second round of digestion and ligation is used to increase resolution [155]. An inverse PCR is then used to capture interactions between the locus of interest and the rest of the genome (one versus all). In the **Hi-C** method, the digested DNA is labeled with biotin, enabling the enrichment for ligation products with streptavidin pull-down [151]. Coupled with high throughput sequencing, Hi-C creates genome-wide contact maps that reflect chromatin organization (all versus all). Further improvements of this technique include in situ Hi-C [153], with an increased percentage of informative ligation products, and single cell Hi-C [156].

Further examination of high-resolution Hi-C data showed that compartments consist of the so-called **topologically associating domains (TADs,** Figure 7C), regions of 100 kb to 1 Mb within which interactions occur more frequently than with adjacent domains [157–159].

Alternative names found in the literature include “insulated neighborhoods” [160], “loop domains” [153] and “contact domains”[161], all of which describe roughly the same structures commonly known as TADs. These domains are separated from each other by boundaries enriched in CTCF binding sites, typically oriented in convergent fashion, which engage in strong interactions that suggest the presence of loops [153]. Boundary CTCF binding sites are often arranged in clusters, forming so-called “super-anchors” [162].

TADs visually appear as squares along the diagonal of the contact matrix, but can be systematically identified by specialized algorithms termed “TAD callers” [163]. The original DomainCaller implemented a directionality index (DI) that measures the interaction bias with upstream (negative DI) or downstream (positive DI) regions [158]. In TAD boundaries, the DI changes its sign to reflect the abrupt change in the polarity of DNA interactions. Initial analyses determined that TADs were remarkably conserved across cell types or even species [153,158]. Nevertheless, increasing Hi-C resolution revealed that TADs are further organized into smaller domains called sub-TADs, which are often cell-type specific and exhibit weaker insulation [153,164]. Accordingly, some TAD callers such as Arrowhead [153] implement multiscale approaches to explore the entire TAD hierarchical structure. Moreover, single cell Hi-C reveal substantial heterogeneity between single cells, implying that the TADs in bulk Hi-C emerge from population averages and not physical structures found in individual cells [165,166].

Contained within TADs are smaller loops of a few kb, often referred to as **regulatory or functional loops**, which mediate the interaction between enhancers and gene promoters [167]. The existence of DNA looping was originally posited in the 1980s as one of several mechanisms [168] to explain previous observations of transcription factors acting at a distance [169,170]. Despite early reports supporting the looping model [171], long-range interactions remained controversial until recently [172]. The advent of 3C technologies allowed the systematic identification of loops and their association with transcription [147]. Critically, these contacts were proven to directly induce transcription in experiments with forced chromatin loops between the *Hbb* promoter and its enhancer [173]. More recently, a new entity termed “chromatin nanodomains” (CNDs) has been proposed as an intermediate step between TADs and loops [174].

3.2.5 Mechanisms of chromatin organization

The observation that TAD boundaries are enriched for **CTCF binding** pointed to a role for this protein in the formation of insulated domains [158]. Initially characterized as a transcription factor, CTCF has long been considered the primary insulator in mammals [175], with some early evidence linking this function to DNA looping [176]. Nevertheless, the insulation function of CTCF may not be entirely dependent on loop extrusion [177]. The vast majority of CTCF motifs at loops anchors are in convergent orientation, suggesting this pattern is a requirement for loop formation [153]. Indeed, the removal or change in orientation of CTCF sites can disturb a TAD boundary, resulting in ectopic contacts between gene promoters and *cis*-regulatory elements that are normally isolated from each other [160,178]. In line with this, loss of TAD

boundaries may cause aberrant expression of oncogenes and developmental factors, leading to cancer and birth defects [179–181]. Nevertheless, since only 15% of CTCF binding sites are located in boundary regions, it soon became clear that the binding of this protein by itself was insufficient to induce the formation of TADs [158].

Subsequent studies revealed that cohesin frequently co-localizes with CTCF at TAD boundaries, but also at the anchors of smaller loops that connect enhancers and promoters [153,164]. **Cohesin** forms a ring-shaped structure containing the subunits SMC1, SMC3, RAD21 and STAG1/2, loaded onto chromatin by NIPBL and MAU2, and unloaded by WAPL and PDS5A/B [182]. Cohesin, like condensin, is a member of the structural maintenance of chromosome (SMC) family, highly conserved ATPases that topologically encircle DNA to produce loops [183]. Originally identified in connection with DNA repair [184], cohesin is involved in sister chromatid cohesion [185,186] and, together with condensin, also in chromatin condensation in preparation for mitosis [185,187]. Several reports in the early 2000s suggested that cohesin and its loading factor NIPBL were implicated in transcriptional regulation, possibly by interfering with enhancer-promoter interactions [188,189]. This function was attributed to transcriptional insulation in cooperation with CTCF, which was necessary for cohesin enrichment at specific loci [190,191]. Shortly afterwards, it was shown that cohesin and CTCF in fact mediate transcriptional activation via long-range interactions between enhancers and promoters [192,193].

Connecting these findings with the observations from Hi-C studies and polymer simulations, the "**loop extrusion model**" was proposed [194–196]. According to this model, chromatin loops are formed by the extruding activity of SMC proteins such as cohesin or condensin, which progressively reel DNA until blocked by a CTCF protein in proper orientation (Figure 8). This process occurs recurrently as loops form, grow and eventually dissociate [194]. It is believed that CTCF stabilizes these loops by creating a physical barrier for cohesin while preventing its unloading by WAPL [197]. The loop extrusion model has been validated by single-molecule imaging showing that cohesin can diffuse rapidly on DNA until it encounters DNA-bound CTCF [198] and, strikingly, real-time imaging of DNA loop extrusion by condensin *in vitro* [199]. In interphase, loop extrusion by cohesin gives rise to TADs and enhancer-promoter contacts, but these features disappear in prophase as cohesin is unloaded and chromatin is compacted by condensin into consecutive loops [200,201]. When cells exit mitosis, they quickly reconstruct their 3D organization, with sub-TADs being formed first, some of which converge into large domains later on following a bottom-up hierarchy [202]. The loop extrusion process is dependent on cohesin's ATPase activity, although loops can be maintained without energy input after their formation [203]. It has been proposed that other mechanisms may contribute to loop formation, namely diffusion by Brownian motion and pushing by RNA pol II [148].

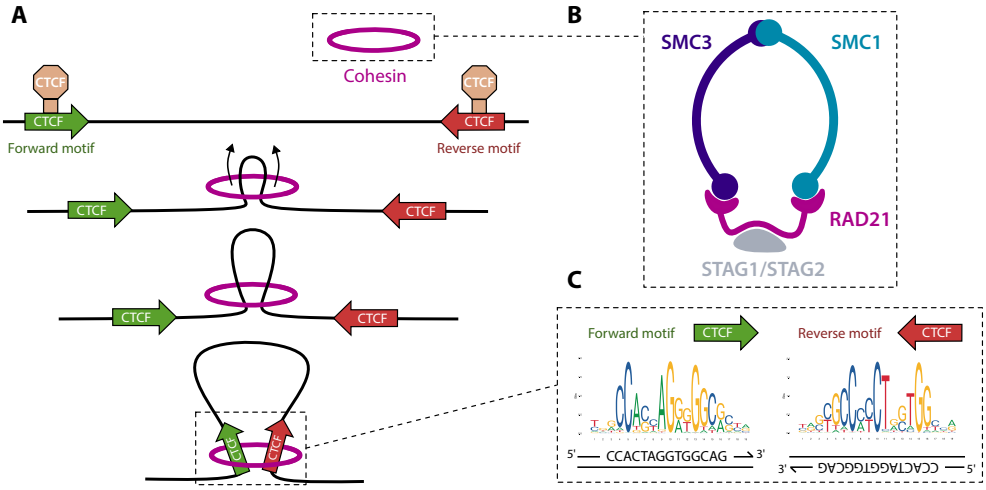


Figure 8. Stabilization of cohesin by convergently-oriented CTCF proteins. (A) According to the loop extrusion model, cohesin reels in DNA until it encounters two convergently oriented CTCF binding sites. Extrusion by human cohesin is symmetrical [197]. (B) Schematic overview of the cohesin complex (C) CTCF motifs at loop anchors are arranged in opposite orientation at each strand.

Although both TADs and enhancer-promoter interactions involve loop extrusion by cohesin, there are differences between these layers of spatial organization. **While CTCF is present at the vast majority of TAD boundaries, it is only found at a small fraction of enhancer-promoter loops** [164]. Moreover, the cohesin complex at TAD boundaries can contain either STAG1 or STAG2 together with CTCF, whereas cell-type-specific contacts are preferentially bound by STAG2, which cannot be replaced by STAG1 [204]. Instead of CTCF, the anchors of these cell-type-specific interactions are frequently occupied by another DNA-binding zinc factor called **YY1** [205,206]. Depletion of YY1 leads to changes in gene expression and loss of enhancer-promoter loops, which are restored upon recovery of YY1 levels. Like CTCF, YY1 is ubiquitously expressed and forms homodimers, suggesting it stabilizes cohesin in an analogous manner. Interestingly, almost 30% of CTCF-binding sites are co-occupied by YY1, particularly at conserved CpG islands, which implies a potentially cooperative action in 3D genome organization [207]. The **Mediator** complex has also been implicated in short-range interactions in collaboration with cohesin [164,193], but recent studies indicate it may act as a functional rather than an architectural bridge between enhancers and promoters [208,209]. Thus, while Mediator is not required for physical contacts, it relays information from transcription factors to RNA pol II, contributing to the assembly of the PIC. In addition, cohesin at non-CTCF sites may be stabilized by other transcription factors [210], such as OCT4 [211].

Consistently with the loop extrusion model, depletion of either CTCF [212] or cohesin [213] results in loss of all CTCF-mediated loops, whereas loss of the cohesin release factor WAPL increases retention of cohesin, leading to longer loops [214]. Interestingly, the changes in

chromatin structure observed in these experiments had small effects on gene expression, with only a few hundred of genes found differentially expressed [212,213]. This suggests additional layers of spatial organization beyond cohesin-mediated loops. Indeed, there are several other mechanisms whereby transcription factors can shape the 3D genome independently of both cohesin and CTCF [215]. For example, **LDB1** is an adaptor protein that dimerizes and forms loops [173] upon recruitment by transcription factors such as GATA1 or TAL-1, as it does not bind DNA directly [216]. Besides, interaction with specific nuclear landmarks such as nuclear pores or the lamina also contributes to genome organization [215].

Nevertheless, the other main driver of chromosomal organization aside from loop extrusion is thought to be **phase separation, which divides chromatin into compartments** [148]. Given its polymeric nature, chromatin folds in such a way that epigenetically similar states (inactive or inactive) cluster together, thus minimizing interactions with other state. Thus, contacts within the active compartment are mediated by interactions between transcription factors and components of the transcriptional machinery. This mechanism is largely uncoupled from loop extrusion, as shown by the fact that compartments remain properly segregated after the removal of CTCF loops [212]. Accordingly, the establishment of compartments in the transition from mitosis to interphase occurs more slowly than the formation of TADs and loops [217]. The two mechanisms, phase separation and loop extrusion, coexist independently, but at the same time influence each other. For example, TADs can bring together regions that can be classified as A and B compartments, which would normally remain spatially segregated [148]. Along these lines, removal of cohesin leads to an increase in compartmentalization [213], whereas depletion of the unloading factor WAPL results in stronger TADs and weaker segregation between A and B compartments [214].

3.2.6 Perturbations of genome structure in disease

Alterations in the 3D organization of the genome lead to aberrant transcriptional regulation that can result in disease or even death. Despite the relatively minor effects of CTCF depletion on gene expression, it is absolutely essential, as *CTCF* knockout embryos are embryonically lethal [218]. Germline *CTCF* mutations are associated with intellectual disability due to the loss of enhancer-promoter interactions for genes involved in cognitive development [219]. Somatic mutations in *CTCF* have been reported in multiple malignancies [220], including lymphoid leukemias [221]. Similarly, haploinsufficiency of *YY1* causes intellectual disability with enhancer dysregulation [222]; while somatic mutations are rare, dysregulation of its expression is common in cancer [223].

Genetic ablation of cohesin results in cell death as a consequence of defects in sister chromatid cohesion during mitosis [185,186], but the importance of gene dysregulation in this context is increasingly recognized. Germline mutations in genes encoding cohesin and its regulatory factors cause disorders collectively known as “cohesinopathies”, among which the Cornelia de Lange syndrome is the most common (with 50-70% of cases due to *NIPBL*

mutations) [224]. Somatic mutations in cohesin subunits have been detected in many cancers [220] and are particularly frequent in myeloid neoplasms [225].

In addition to alterations in the genes encoding for cohesin-related proteins and CTCF, which have widespread effects, diseases have been linked to local alterations. Genomic locations co-occupied by both CTCF and cohesin are mutation hotspots in cancer [226], which are associated with chromosomal instability and changes in gene expression [227]. Given the sensitivity of CTCF binding to methylation (further described in section 3.4), hypermethylation of CTCF binding sites can also disrupt TAD boundaries and deregulate gene expression [228]. A recent pan-cancer study of CTCF motifs in TAD boundaries confirmed the presence of frequent somatic mutations and hypermethylation [229].

Moreover, structural rearrangements often affect TAD boundaries both in development [179,180] and cancer [230], leading to aberrant expression of nearby genes. The high frequency of chromosomal abnormalities at TAD boundaries may in fact be an untoward consequence of the machinery required to maintain genome organization, as topoisomerase 2B (TOP2B) accumulates at loop anchors and induces double strand breaks to untangle DNA during extrusion and prevent torsional stress [231]. The illegitimate repair of these TOP2B-induced double strand breaks can occasionally join two regions that are proximal in space but linearly distant, resulting in chromosomal translocations [232]. Indeed, breakpoints of translocations recurrent in cancer, such as *MLL* rearrangements, are enriched at loop anchors [231]. This could also explain why cohesin mutations are largely mutually exclusive with chromosomal aberrations in AML.

3.3 The role of histones in gene regulation

Histone proteins consist of a well-ordered globular core ("histone fold") flanked by intrinsically disordered tail domains ("histone tails") [233]. In the nucleosome, the globular domains, composed mainly of basic residues, form stable dimers with each other to constitute the histone octamer – these account for the majority of DNA-histone interactions [234]. The N-terminal tails of the four core histones and the C-terminal tail of H2A protrude from the NCP and harbor abundant arginine and lysine residues, which confer them a positive charge [235]. This charge enables interaction with negatively charged DNA, which stabilizes DNA wrapping in the nucleosome and facilitates the formation of higher order chromatin structures [236]. Moreover, the tail of H4 (and other histones to a lesser extent) contributes to inter- and intra-nucleosome interactions by binding the acidic patch of H2A/H2B, a region enriched in acidic residues [234,236]. However, histone tails can be deleted without major effects on nucleosome integrity, indicating that they are not essential [237].

The tail domains contain a large number of sites that can be target of post-transcriptional modifications (PTMs), which modulate the charge of the tail and thus alter the electrostatic interactions supporting chromatin structure [235]. In addition to this direct effect, PTMs recruit transcription factors and remodelers to indirectly regulate chromatin structure. For this

reason, it has been proposed that PTMs collectively create a "histone code" that can be read or written by other proteins [238,239]. Besides, histone variants also modulate the stability of the nucleosome and play a key role in transcription [234].

3.3.1 The histone code

The existence of histone tail PTMs has been known since 1964, when Vincent Allfrey showed that acetylation and methylation are incorporated after synthesis of the polypeptide chain [240]. Crucially, he also demonstrated that acetylation promotes RNA synthesis by relieving the repressive effect of histones on transcription. Since then, several other histone modifications have been identified, especially through the use of mass spectrometry in the last two decades [241]. However, acetylation and methylation remain the best understood, followed by phosphorylation and ubiquitination [242]. Furthermore, modifications of the central globular domain have also been reported to influence nucleosome stability [243]. Finally, there is increasing evidence for a role of modifications of the terminal tip of the histone tails (as opposed to the better characterized PTMs of the side chain) [244].

In their seminal article, C. David Allis and Brian Strahl put forward the **"histone code"** as a language encoded on histone tail domains that could be read, written or erased by specific proteins [238]. A stepping stone for this hypothesis was the realization that bromodomain-containing proteins bind acetylated lysines, constituting the first example of **"reader"** [245]. The existence of "**writers"** that add histone modifications had been known since 1996, with the identification of a histone acetyl transferase (HAT) in *Tetrahymena* and its yeast homolog Gcn5p [246,247]. Finally, that same year also saw the isolation of a histone deacetylase (HDAC) [248], an "**eraser**" that removes acetylation, although the existence of such enzymes had been previously reported [249]. According to this model, histone modification is therefore a dynamic process regulated by groups of enzymes with opposing activities.

The histone code hypothesis predicted that similar mechanisms would be identified for other modifications, and that was indeed the case. Around the time of its publication, enzymes involved in histone phosphorylation [250], dephosphorylation [251] and methylation [252,253] were identified, shortly followed by the first lysine demethylase (LSD) in 2004 [254]. Earlier studies had suggested the existence of histone demethylases [255], but their molecular identity had remained elusive for decades. Thus, the identification of LSDs constituted the missing proof that all histone modifications can be written or erased by specific enzymes (Table 1), as claimed by the "histone code" hypothesis [256]. This framework can be further broadened to a general "**epigenetic code**" by considering higher order chromatin and DNA methylation, which follows similar rules involving reading (CpG-binding proteins), writing (DNA methyltransferases) and erasing (DNA demethylation via oxidation) [239,257]. Additional elements of this language have been proposed recently, including the "ink" represented by metabolites or the "paper" symbolized by histone variants and chromatin remodelers [258].

Table 1. Common histone modifications and their associated enzymes. This table has been compiled with information from multiple sources, listed in the last column. A comprehensive catalog of histone modifications is available at [259].

Modification	Residues	Writer	Eraser	Readers	Source
Acetylation	K-ac	Acetyltransferases (HAT)	Deacetylases (HDAC)	Bromodomain, PHD finger	[260]
Methylation	K-me1/2/3 R-me1/2/3	Lysine methyl-transferase (KMT) Protein arginine methyl-transferases (PRMT)	Lysine-specific demethylases (KDM) Jumonji C-containing proteins (JmJC)	ADD, Tudor, WD40, PHD, MBT, PWWP	[261,262]
Phosphorylation	S-ph, Y-ph T-ph, H-ph	Multiple kinases	Multiple phosphatases	14-3-3 BIR, Tandem BRCT	[263]
Ubiquiylation	K-ub	Ubiquitin ligases (RNF family)	Deubiquitinating enzymes (DUB)	RNF168, RAP80, 53BP1	[264]
Sumoylation	K-su	SUMO ligases	SUMO-specific proteases	SIM-containing proteins	[265]
ADP ribosylation (mono, poly)	E-ar, R-ar K-ar, S-ar	Poly(ADP-ribose) polymerase (PARP), sirtuins	ADP-ribosylhydrolases (ARHs)	Macrodomains, PBZ, WWE, PBM	[266]
GlycNAcylation	T-og, Ser-og	O-GlcNAc transferase (OGT)	O-GlcNAcase (OGA)	14-3-3	[267,268]
Citrullination (deimination)	R > Cit	Peptidyl arginine deiminases (PADs)	None known	None known	[269]
Crotonylation	K-cro	Crotonyltransferases (HCT)	Decrotonylases (HDCR)	DPF domain, YEATS domain, bromodomain	[270]

Another key prediction of the histone code was that histone tail modifications may be interedependent and act in combination [238]. Identification of “bivalent domains” confirmed that overlapping histone marks carry a different signal from that of those marks in isolation [271]. In this case, trymethylation of the lysine 27 in H3 (H3K4me3) is associated with active promoters and H3K27me3 is associated with repressive chromatin states, but regulatory elements marked by both are in an intermediate “poised” state. In enhancers, H3K4me1 precedes H3K27ac deposition, priming enhancers for further activation as well as providing a molecular “memory” of prior activation [272,273]. Several more examples of combinatorial effects have been described [141].

Combinations of chromatin marks can be integrated into so-called **chromatin states**, which precisely delineate functionally distinct genomic regions such as promoters or enhancers [274]. These inferred functional associations are a result of specific recognition by reader proteins that contain motifs able to distinguish residues based on their methylated stated and surrounding sequence. These motifs include PHD, WD40 and Tudor domains capable of binding both arginine and lysine residues [275,276], as well as several others restricted to lysines. Interestingly, the same histone modifications can also be associated with opposing activities depending on which proteins recognize them, and thus their context [277].

Given the central role of histone modifications in the regulation of gene expression, it is hardly surprising that defects in their addition, removal or interpretation are intimately

linked to cancer development [258]. Accordingly, a large number of emerging therapies are targeted at epigenetic readers, writers and erasers [278].

3.3.2 Exploration of the histone modification landscape

One of the main tools to investigate the patterns of histone modifications is **Chromatin Immunoprecipitation (ChIP)** [279]. Originally described in the 1980s [280,281], this technology can be applied to identify *in vivo* binding of any protein associated with chromatin, including histones, but also transcription factors or RNA polymerases. These proteins are first crosslinked to DNA, forming covalent bonds, by treating cells with formaldehyde [282]. This is followed by fragmentation of the fixed material, usually by sonication, and immunoprecipitation of the DNA-bound protein of interest with specific antibodies. Finally, the DNA is released from crosslinked proteins and purified for further analysis, which can include a variety of molecular biology techniques. In early versions of this method, binding regions were detected by dot blot or Southern blot, which made it possible to establish that histones remain bound to DNA during transcription [280]. Eventually, it was coupled with quantitative PCR (qPCR) to measure enrichment at specific regions [283].

A major improvement was the ability to assess regions across entire genomes by hybridizing the enriched DNA with microarrays, a method called ChIP-on-chip [284]. However, this assay had important limitations: coverage and resolution were restricted by the features included in arrays and constraints imposed by hybridization chemistry, which also resulted in noisier signal due to cross-hybridization. Thus, it was promptly superseded by **ChIP-seq**, which combined ChIP with next-generation sequencing to achieve whole genome coverage at high resolution [285]. Initial studies using these technology mapped chromatin states genome-wide and determined how different histone marks are related to gene expression [286,287]. In the ensuing years, large consortia such as ENCODE [288] or Blueprint [289] capitalized on ChIP-seq to define epigenetic states across different tissues.

Despite its popularity, ChIP-seq requires large cell numbers, which precludes the identification of phenomena restricted to small subpopulations. To overcome this drawback, single cell approaches have been developed [290]. Another hurdle is the presence of artifacts, often as a result of cross-linking and pre-amplification by PCR, among other factors [291,292]. While native ChIP-seq (N-ChIP) avoids the use of crosslinking [293], it is limited to proteins with high stability and it may be affected by chromatin rearrangement during the process [294]. A promising alternative is **CUT&RUN**, which requires a small amounts of starting material and provides a high signal-to-noise ratio at a lower sequencing depth [295]. Briefly, CUT&RUN tethers a protein A/micrococcal nuclease (pA-MNase) fusion protein to antibody-labelled protein loci for directed cleavage. Its successor CUT&Tag follows a similar strategy, but using hyperactive Tn5 transposase pre-loaded with sequencing adapters instead, resulting in lower costs and facilitating its use in single cell platforms [296].

3.3.3 Histone acetylation and methylation in transcription

Histone acetylation

Histone acetylation was the first histone modification to be described, as early as 1963 [297]. Shortly afterwards, Allfrey described its association with active transcription [240]. A long-held view was that this link was a result of the physical properties of acetyl groups [298], which neutralize the positive charge of lysine residues, thereby decreasing their affinity for DNA [299]. This effect results in destabilization of the nucleosome becomes and increased accessibility of the DNA to transcription factors [300]. Although this direct (*cis*) mechanism is certainly crucial, the discovery of bromodomains and a slew of proteins containing them has revealed that histone acetylation also mediates indirect (or *trans*) effects as a docking site for reader proteins [256]. As mentioned before, bromodomain-containing SWI/SNF binds acetylated regions [143], providing an example of indirect mechanism for chromatin accessibility regulation by acetylation. The human genome encodes 42 bromodomain-containing proteins that harbor a total of 56 distinct bromodomains [260]. Notably, the bromodomain and extra-terminal (BET) family member BRD4, a critical mediator of transcription, is an attractive pharmacological target in cancer [301].

The addition of acetyl marks is catalyzed by HATs, which are classified into Type A if they have nuclear activity (such as HAT1) and Type B if they reside in the cytoplasm [302]. Type A HATs transfer the acetyl group after nucleosome formation, whereas type B enzymes modify free histones before their deposition. Type B enzymes can be further subdivided into five subfamilies in mammals, namely GNAT (KAT2A/KAT2B), MYST (MOZ, MOF, TIP60), p300/CBP, basal TFs (TAF1) and nuclear receptor cofactors (SRC1). The opposite action of HATs is mediated by HDACs, which are classified into three zinc-dependent classes (I, II and IV) and a NAD+-dependent class (III) consisting of sirtuin proteins [302].

Multiple lysines in histone tails can be subjected to acetylation, each of which is preferentially recognized by specific readers, writers and erasers [260]. This specificity argues against a model in which the only function of acetylation is charge neutralization, despite early experiments showing a redundant role of different lysines in transcription [303]. Accordingly, specific associations between certain acetylated positions and regulatory elements or gene expression have been reported [304,305]. Most notably, **H3K27ac, which is deposited by CBP (KAT3A) or p300 (KAT3B**), is associated with active enhancer and promoter elements [306]. Similarly, the combination of H3K9ac and H3K14ac is frequently present at CREs together with H3K27ac, whereas H3K14ac alone is enriched at a subset of inactive promoters marked by H3K27me3 [307]. More recently, acetylation of lysines in the H3 globular domains (H3K64ac and H3K122ac) [308] or the H4 tail acetylation H4K16ac [309] have been linked to subsets of enhancers not marked by H3K27ac. Contrary to PTMs in the H3 tail, which mostly depend on specific readers to modulate transcriptions, these three modifications directly affect chromatin structure by interfering with nucleosome stability or inter-nucleosomal interactions [310–312].

Histone methylation

Another prominent **histone modification is methylation**, which can occur on all basic residues: arginines, lysines and histidines [277]. Discovered in 1964 [313], lysine methylation has been extensively studied, but the effect of arginine methylation has only recently started to become clear [262]. Histidine methylation is very rare and its relevance remains unknown. Lysines can be mono- (me1), di- (me2) or trimethylated (me3), whereas arginines can be monomethylated (me1), symmetrically methylated (me2s) or asymmetrically methylated (me2a) [277]. In contrast with acetylation, which is clearly correlated with active transcription, the effect of methylation depends on the residue and number of methyl groups. Thus, the monomethylations of H3K27, H3K9, H4K20, H3K79 are enriched in actively transcribed genes, whereas trimethylation of H3K27, H3K9, and H3K79 marks repressed regions [286].

Remarkably, **H3K4me3** is highly enriched at active promoter regions and **H3K4me1** is linked to enhancer function [314], whereas heterochromatin is enriched for **H3K27me3** [286] (facultative) and **H3K9me3** (constitutive) [315]. While H3K36me3 is generally found at transcribed gene bodies, it may also contribute to repression [316]. Some of these marks, like H3K27me3 or H3K9me3, can be found in broad domains, often containing repressed genes developmental genes [261].

Methylation is deposited and removed by enzymes with high specificity for certain positions and degrees of methylation [261,275]. Methyltransferases catalyze the donation of methyl groups from S-adenosylmethionine to histones and can be grouped into three families [277]. Most lysine methyltransferases (KMT) contain a SET domain, including KMT1A, the first KMT ever discovered [252]. KMT1A specifically metylates H3K9 (H3K9me3) from a monomethylated state (H3K9me1), and similar degrees of selectivity can be observed in other enzymes of the family. A special group of proteins within this category is the **Polycomb repressive complex 2 (PRC2), which establishes H3K27me3** (Box 2) [317]. The second class of KMTs is represented only by DOT1L and exclusively methylates H3K79 [318,319]. On the other hand, arginines are methylated by the multiple members of the protein arginine N-methyltransferase (PRMT) family [320].

Box 2. Polycomb and trithorax proteins

The **polycomb group (PcG)** of proteins were originally identified in Drosophila as repressors of homeotic genes [321]. Thritorax proteins (TrxG) were discovered shortly afterwards as anti-silencers that activated the expression of homeotic genes [322,323]. Both affect histone methylation, but the PcG complex PRC2 deposits H3K27me3 to repress transcription, whereas TrxG proteins are KMTs that trimethylate H3K4 to induce transcription [317]. In humans, PRC2 consists of the catalytically active EZH2 and the accompanying subunits EED and SUZ12. The PRC1 complex acts as a “reader” by binding H3K27me3-marked regions and mediating silencing.

Removal of lysine methylation is carried out by lysine demethylases (KDM), the first of which was discovered in 2004 after long decades of debate regarding their existence [254]. KDM1A, also known as LSD1, contains an amine oxidase domain that demethylates H3K4me2/1. Another class of KDMs harbors JmjC domains which catalyze the oxidation of methyl groups, as originally shown for the H3K36me2 demethylase, KDM2A [324]. The existence of arginine demethylases (RDM) remains controversial, but it has been shown that certain JmjC-containing KDMs exhibit arginine demethylation activity in vitro [262].

Despite the strong association between histone PTMs and gene expression, their role in transcriptional regulation may be more limited than originally thought [92]. Experiments in drosophila revealed that transcriptional regulation can occur in the absence of H3K4 methylation [325] and that point mutations in H3K27 lead to a loss of PRC-mediated repression, suggesting that acetylation mainly antagonizes H3K27me3 [326]. Similarly, H3K4me1 in enhancers seems to be dispensable for eRNA synthesis and transcription from promoters [327].

3.4 DNA methylation

In humans and other vertebrates, **DNA methylation commonly refers to the covalent addition of a methyl group to the position 5 of cytosine in a CpG context**, yielding 5-methylcytosine (5mC) [328]. These methyl groups project into the major groove of DNA and modify the functional state of regulatory regions, without affecting the DNA sequence. The discovery of 5mC dates back from 1925, as one of the hydrolysis products of *Bacillus tuberculosis* nucleic acids [329], though it was not confirmed in mammalian DNA until 1945 [330]. A role for DNA methylation in transcription and differentiation was proposed by several authors in the early 1970s [331,332], and subsequently formalized and expanded by two papers published in 1975 [333,334]. In these landmark publications, Riggs and Holliday & Pugh independently hypothesized that DNA methylation silences transcription in differentiation by affecting the binding of regulatory proteins, and this epigenetic mark is inherited upon cell division. Experimental support for this theoretical model was gathered in the following years, firmly establishing the link between methylation and gene expression by 1980 [335].

Given that 1% of the human genome consists of 5mC, it has been sometimes referred to as “the fifth base” [336]. It is estimated that roughly 70-80% of the CpG sites are methylated in somatic cells, although the exact percentage greatly varies across tissues [328,337]. However, methylated cytosines can be converted to thymine by spontaneous or enzymatic deamination [338,339], leading to an underrepresentation of the CpG dinucleotide of 20% of what would be expected [340]. The exception to this globally methylated and CpG-poor landscape are the so-called **CpG islands** (CGI), regions of 1000 bp on average characterized by high GC content, little CpG depletion and absence of methylation [341].

DNA methylation is a key element of lineage specification. Genome-wide studies have identified tissue-specific differentially methylated regions (DMRs), with tissues or cell types belonging to the same organs clustering together [342,343]. Most of these regions are

hypomethylated and overlap with transcription factor binding sites (TFBS) and *cis* regulatory regions related to tissue-specific functions.

3.4.1 Cellular functions of DNA methylation

DNA methylation has been associated with **transcriptional repression** since the 1970s, but studies in the last decade have drawn a more nuanced picture [344,345]. For example, methylation in the gene bodies positively correlates with active transcription [346,347]. Even so, the traditional view generally applies to promoters, whose methylation is inversely correlated with gene expression [348]. This is particularly evident in the context of X chromosome inactivation [349] and imprinting [350], two key biological processes in mammalian development. The causal relationship between methylation and expression involves direct and indirect mechanisms:

- **Impaired TF binding:** DNA methylation can directly repress transcription by preventing the binding of activating transcription factors to regulatory regions, as is the case of MYC [351] or the ETS family [352]. However, other TFs can bind methylated regions, such as SP1 [353] or YY1 [352]. In fact, an unbiased screen determined that the binding of 23% of TFs is inhibited by methylation, but it was enhanced in 34% [354]. This same study concluded that the negative impact of methylation was due to steric hindrance. However, the causality between these two phenomena is not always straightforward – while methylation may dictate TF binding, the methylation status of a genomic region can also be affected by the presence of TFs [355].
- **Recruitment of repressors:** methylated CpGs can be recognized by "reader" methyl-binding proteins (MBPs), which indirectly contribute to repression via different mechanisms [356]. These proteins belong to three structural families: the MBD family (e.g. MBD1, MBD2), most of whose members recognize CpG-rich methylated in a non-sequence specific manner [357,358]; the zinc finger family (e.g. ZBTB33 or ZBTB4), with the ability to bind either single or double CpGs [359]; and the SRA family (UHRF1 and UHRF2), which recognize hemimethylated sites and recruit DNA methyltransferases and HDAC1, which leads to transcriptional silencing [360,361].

Besides their role in transcriptional repression, MBPs have other functions such as DNA repair (MBD4 [362]) or methylation maintenance (UHRF1 [361]). Although the mechanism is not fully understood, MBD2 and MBD3 may also be implicated in gene activation [362], which may explain the positive correlation of methylation and transcription at gene bodies. In addition, as described above, a plethora of TFs have shown at least some ability to bind methylated sequences. Aside from gene activation, possible consequences include the opening of chromatin thanks to pioneer TFs and splicing regulation [355].

DNA methylation may weaken the binding of CTCF [363,364], which plays critical roles as insulator, transcriptional repressor or activator and architectural protein [175]. Thus, aberrant methylation can disrupt CTCF-dependent TAD boundaries, resulting in dysregulated

expression of neighboring genes [365]. A notable example is the overexpression of *PDGFRA* in glioma with *IDH1/2* mutations, as a result of TAD fusion and enhancer hijacking [228]. A genome-wide study revealed that 41% of variable CTCF binding across cell types can be attributed to differential methylation, with a 87% average reduction in occupancy in methylated regions [366]. However, 36% of the variable CTCF sites tested did not overlap with variable methylation, and the remaining 23% were insensitive to methylation. This is in line with the observation that the demethylating agent decitabine only altered CTCF occupancy in a fraction of the genome [367]. A possible explanation is that not all CTCF motifs contain a CpG in position 2, whose methylation inhibits interaction with zinc finger 7 of CTCF [368]. Intriguingly, CTCF binding itself can initiate local demethylation in CpG-poor regions [369].

3.4.2 Writers and erasers of DNA methylation

Methylation occurs after DNA synthesis by methyl transfer from S-adenosylmethionine (SAM) to cytosine [335]. *De novo* methylation is **established by the methyltransferases** ("writers") DNMT3A and DNMT3B in combination with DNMT3L, which is catalytically inactive, but stimulates the activity of DNMT3A and DNMT3B [370,371]. There are differences in the function and specificity of these two enzymes owing to differences in the N-terminal domain [372] – while DNMT3A targets short interspersed repeats, DNMT3B is specific for satellite repeats [373]. Two isoforms of DNMT3A and more than 30 isoforms of DNMT3B have been reported, with different expression patterns and functions [374]. The accessory factor DNMT3L is only expressed in the germ line and the early embryo [375] and is essential for genomic imprinting together with DNMT3A [370,376]. However, it is dispensable for DNMT3B-mediated methylation, which could be explained by the finding that DNMT3L preferentially stabilizes DNMT3A [377]. These mechanisms are summarized in Figure 9.

These methylation patterns must be preserved across generations to ensure that distinct cell types retain their identity. However, upon DNA replicated yields hemi-methylated DNA – only the parental strand is methylated, whereas the newly synthesized strand is unmethylated [328]. Thus, Holliday and Pugh predicted the existence of **maintenance methyltransferases** that recognize hemi-methylated DNA [333]. Isolated in 1983 by Bestor and Ingram [378], DNMT1 is the main enzyme responsible for methylation maintenance, and its loss results in embryonic lethality [379]. The activity of DNMT1 depends on the interaction with a number of accessory factors [380]. UHRF1 favors the recruitment of DNMT1 to hemimethylated CpG sites [361], whereas PCNA facilitates the action of DNMT1 on newly synthesized DNA [381]. Although it was initially thought that DNMT1 was sufficient to maintain methylation patterns, DNMT3A and DNMT3B also participate in this process, in which they compensate for inefficient activity of DNMT1 at certain locations, such as repetitive elements [382,383]. Cooperation between DNMT3B and DNMT1 has also been observed in cancer cells, where they maintain silencing of tumor suppressor genes [384]. Conversely, DNMT1 can also be involved in *de novo* methylation in cooperation with the DNMT3 enzymes [385].

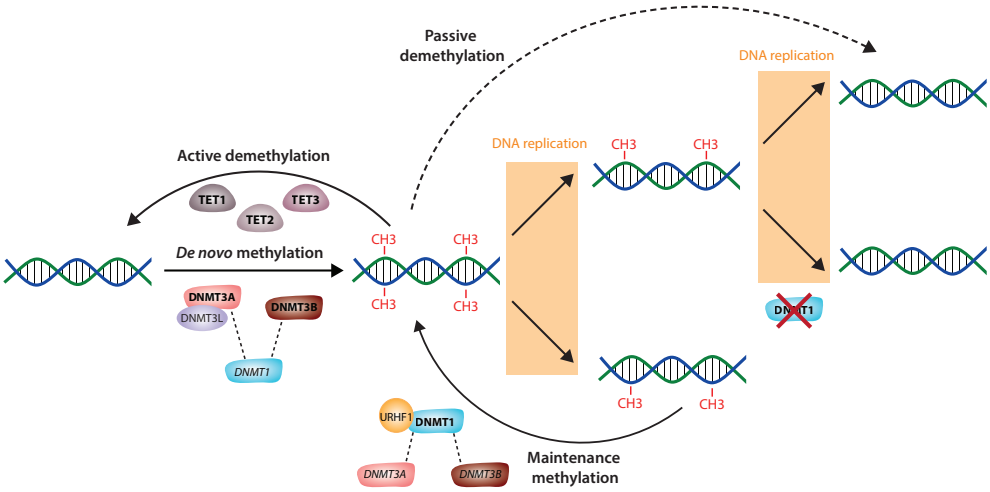


Figure 9. Mechanisms and enzymes regulating DNA methylation. The addition and maintenance of methyl groups is mediated by proteins in the DNMT family, whereas demethylation is either catalyzed by TET enzymes or a result of replication without maintenance. The key enzyme in each step is written in bold, accessory enzymes are in italics.

In the course of differentiation, DNA methylation must be removed to establish a permissive state for gene expression [345]. In human development, the two waves of *de novo* methylation are followed by corresponding waves of demethylation. The first occurs after fertilization, in which the embryo loses gamete-specific DNA methylation to enable pluripotency; the second takes place in primordial germ cells (PGCs) and allows sex-specific imprinting in later stages [344]. These two waves require a combination of **passive and active demethylation**. Passive demethylation is achieved by replication-coupled dilution of 5mC in the absence of DNMT1, although small amounts of DNMT1 can preserve imprinting in embryogenesis [386]. However, the extent of 5mc loss after fertilization cannot be explained by passive dilution alone, suggesting the need for an active enzymes ("erasers") that remove 5mC independently of replication [387]. Although different mechanisms had been proposed [388], it was not until recently that TET1 [389] and its homologs TET2 and TET3 [387] were unambiguously identified as critical mediators of this process.

Proteins of the **TET family** mediate the oxidation of 5mC to 5-hydroxymethylcytosine (5hmC), which is not recognized by DNMT1 and thus passively lost upon replication [390]. Furthermore, 5hmC can be further converted to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC), which can be removed by thymine DNA glycosylase (TDG) coupled with base excision repair (BER) or passively lost as well [391–393]. There are differences in the structure, activity and expression patterns of the TET proteins, indicating they are not redundant [394,395]. All three are expressed at different stages of embryogenesis and gametogenesis, but only TET2 and TET3 remain expressed in adult tissues, including the hematopoietic system [394,395]. However, at least in certain contexts, loss of one protein can be partially compensated

by other members of the family. For example, deficiency of TET1 or TET2, individually or combined, results in abnormal methylation but does not compromise embryo viability, indicating a compensation by TET3 [396–398]. Similarly, while TET2 is particularly important for hematopoiesis [398], its loss can be compensated by TET3 [399]. Although controversial, another mechanism involving the conversion of 5mC to thymine by activation induced deaminase (AID) has been proposed [388]. As described for TET, this thymine would be excised by TDG and subsequently repaired by the BER pathway.

In somatic tissues, **differentiating cells undergo microwaves of *de novo* methylation and demethylation** that lead to widespread discrepancies in DNA methylation between tissues [344]. Consistently with the notion that mature cells exhibit lower methylation levels than their progenitors, active demethylation allows access of TF at regulatory regions controlling tissue identity [390]. In particular, TET2 binds to cell-type specific enhancers and reshapes their TF accessibility [400]. The participation of hypermethylation in differentiation is less well understood, but *de novo* methylation by DNMT3A is required in neurogenesis [401] and hematopoiesis [402]. It has been proposed that a DNMT3B isoform could replace DNMT3L as an accessory factor and recruit DNMT3A in somatic cells [403]. On the other hand, deletion of both DNMT3A and DNMT3B in B cells does not impair maturation, but modulates B cell activation [404]. All in all, while active demethylation by *TET2* seems to be a common mechanism in differentiation, DNMT3A/B may be restricted to certain stages or tissues.

The regulation of **DNA methylation is tightly coupled with the histone code**, suggesting an interplay between these two epigenetic layers [272]. The ADD domain of DNMT3A/B and DNMT3L preferentially binds to unmethylated H3K4, guiding *de novo* methylation, but is repelled by H3K4me3 [405,406]. In addition, DNMT3L enhances methylation by DNMT3A/B at gene bodies, yet counteracts their activity at CGIs with H3K27me3 in ESCs in order to preserve hypomethylation [407]. On the contrary, the PWWP domain of DNMT3A selectively binds H3K36me3, associated with transcriptional elongation, leading to enriched DNA methylation at actively transcribed genes [346,408]. Similarly, UHRF1 –often found in a complex with DNMT1–exhibits preference for methylated H3K9 positions [409]. Besides, DNA methyltransferases interact with several histone modifiers. For example, the PRC2 subunit EZH2 recruits DNMTs to target promoters [410] and the histone H3K9 methyltransferase SETDB1 interacts with DNMT3A/B, but not DNMT1, at promoter regions [411]. Furthermore, both DNMT1 [412] and DNMT3A [413] associate with HDAC1 to generate a silent chromatin state. Altogether, these observations indicate that DNA methylation may act as a "lock" to stabilize silencing by other mechanisms [345].

3.4.3 CpG islands and open seas

Originally described by Adrian Bird in 1986 [414], CGIs were formally defined by Gardiner-Garden and Frommer as regions of at least 200 bp with average GC above 50% and observed/expected CpG ratio above 0.6 [415]. Using this definition, the Human Genome Project

identified 28,890 CpG islands in non-repetitive regions of the genome [340]. Subsequent *in silico* and *in vitro* analyses of the human genome have yielded between 24,000 and 27,000 CGIs [341,416]. Given that any definition is necessarily arbitrary, a single correct estimate cannot be unequivocally determined, but it is likely to be within this range. Besides, Irizarry and colleagues recently defined "CpG shores" as 2 kb regions around CGIs where most tissue-specific methylation takes place [417]. This definition was further amended (Figure 10) to include "CpG shelves" as the 2-kb regions flanking CpG shores [418] and "open sea" as any region containing isolated CpGs in the rest of the genome [419].

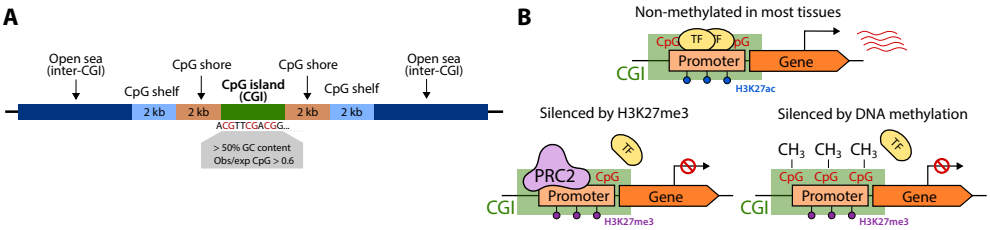


Figure 10. Definition of CpG islands and their involvement in gene regulation. (A) Annotation of CpG islands, shores, shelves and open seas, adapted from [420]. (B) Promoters in CpG islands are frequently in an active state, but they can be repressed by either H3K27me3 or methylation during differentiation.

Around 70% of annotated gene promoters are associated with a CGI, including most housekeeping genes and a few genes involved in development and differentiation [421]. These CGI promoters are typically in a non-methylated state, even when the corresponding gene is inactive [348]. Furthermore, CGIs exhibit features associated with active transcription: open chromatin, H3 acetylation and H3K4 methylation [287,348,422]. Another distinct feature of CGI promoters is that they are frequently bound by CTCF, which may play a role in their 3D organization and activation [423].

The high GC content seems to impair formation of stable nucleosomes, which facilitates transcriptional induction without a requirement for chromatin remodelers [424]. However, **gene expression is silenced when CGI promoters become methylated**, which normally occurs a consequence of chromosome X inactivation and gene imprinting [348]. Even so, **CGI promoters are mainly repressed by the polycomb** complexes PRC1 and PRC2, a more plastic mechanism that facilitates reactivation in development [344]. Indeed, while H3K27me3 is virtually absent in non-CGI promoters, 20% of CGI promoters in ES cells exhibit this mark, half of which lose it upon commitment [287]. The regions that retain H3K27me3 become frequently methylated in differentiation [425], which is made possible by loss of the activating mark H3K4me3. This suggests **affinity of DNA methylation for H3K27me3**, possibly due to interaction between DNMT3A/B and EZH2, and is further evidence that DNA methylation locks prior silencing by other mechanisms during development. Even so, the proportion of methylated CGIs remains relatively small in somatic tissues [341,348].

Approximately half of mammalian CGIs are not associated with any known promoter and thus considered "orphans" [426]. They are often located in intragenic regions and frequently exhibit tissue-specific methylation [427]. Nonetheless, they have the characteristics of functional promoters, including H3K4me3 enrichment, RNA poll II recruitment and presence of Cap Analysis of Gene Expression (CAGE) [426,427]. This indicates that orphan CGIs are tissue-specific alternative promoters subjected to regulation. Besides, orphan CGIs located at poised enhancers can tether them to distantly located genes with CGI promoters, thus acting as determinants of enhancer-promoter compatibility [428].

CGIs retain their high GC content by **remaining unmethylated in the germ line**, thereby avoiding deamination [429]. It has been proposed that this is achieved thanks to the elevated levels of H3K4me3, which protects against *de novo* methylation by DNMT3A [405]. Furthermore, TET1 preferentially binds CpG-rich regions and catalyzes their demethylation [430] while favoring the recruitment of PRC2 and the establishment of H3K27me3 [431]. Regions occupied by PRC2 are protected from methylation by recruiting FBXL10 [432] and sequestering of DNMT3L [407]. Nevertheless, a small fraction of CGIs become methylated in oocytes and sperm, progressively increasing with gamete maturation [433].

Although less than 10% of the CpGs are located in CGIs, the role of methylation in the rest of the genome is not fully elucidated [434]. In gene bodies, methylation of the first exon is associated with transcriptional silencing [435], whereas methylation of the rest of the exons inversely correlates with transcription [346,347]. Contrary to CGI promoters, the majority of **non-CGI promoters are methylated** [348,436] and compacted in stable nucleosomes that are only induced upon binding selective chromatin remodeling [424]. However, the role of methylation at these locations is controversial. While some studies indicated that the activity of non-CGI promoters is independent of methylation [348], others showed negative correlations between methylation and expression [436,437]. Careful analysis of a few loci revealed that methylation of tissue-specific non-CGI promoters leads to silencing [438], supporting the notion that promoter methylation generally represses transcription.

3.4.4 Methylation alterations in cancer

Albeit rare in normal development, many **CGIs become hypermethylated in cancer**, leading to the repression of tumor suppressor genes [439]. A prime example of this phenomenon is colorectal cancer, in which genome-wide aberrant methylation of CGIs defines the "CpG island methylator phenotype" (CIMP) group [440]. This frequently results in repression of the *MLH1* gene, which in turn results in mismatch repair deficiency and microsatellite instability [441,442]. Similar entities have been identified in AML, leading to inactivation of hematopoietic transcription factors such as *CEBPA* [443–445]. As in differentiation, tumor-associated CGI methylation preferentially occurs at regions marked by H3K27me3 [446,447]. Therefore, cell chromatin patterns of cancer stem cells can make them susceptible to aberrant methylation of genes involved in differentiation. Aberrant hypermethylation is strongly

associated with the loss of H3K4me3, which confers protection against DNA methylation [448]. Moreover, sequence features may affect the propensity of CGIs to be methylated, as shown in experiments involving overexpression of DNMT1 [449]. Certain oncoproteins, such as PML-RARA in AML, can also recruit DNMT1 and DNMT3A to CGI promoters and promote hypermethylation [450].

In contrast with CGIs, **the rest of the genome tends to be hypomethylated in cancer** compared to normal tissues, often coexisting with promoter hypermethylation [451,452]. Highly repetitive elements account for most of this global hypomethylation, including LINEs, SINEs, subtelomeric repeats and segmental duplications [453,454]. Induction of hypomethylation in animal models with demethylating agents [455] or via *DNMT1* downregulation [456,457] leads to carcinogenesis, establishing a causal role for hypomethylation in some cancers. Possible mechanisms include increased chromosomal instability, activation of transponsable elements normally silenced by methylation, sequestration of TFs by accessible repeats, alterations in chromatin structure and aberrant expression of noncoding RNAs [458]. Alternatively, hypomethylation and consequent overexpression of specific oncogenes such as *RAS* [459] can promote tumor development. In some cases, hypomethylation causes **loss of imprinting (LOI)** of genes whose expression is normally restricted to one allele, such as *IGF2*. Following its original description in Wilm's tumors [460], LOI of *IGF2* has been reported in a wide range of cancers, including leukemia [461,462]. Experiments in animal models have confirmed that altered expression of genes with LOI induces tumorigenesis and may be an initiating even in cancer [463].

The primary cause of altered methylation patterns in cancer remains elusive, but in some cases it can be linked to mutations in enzymes regulating DNA methylation. For example, *DNMT3A* mutations are present in 20% of AML patients [464], as well as in other hematopoietic malignancies such as myelodisplastic syndromes (MDS) and acute lymphocytic leukemia (ALL) [465]. *DNMT3B* is overexpressed in colorectal cancer [466] and glioma [467], and it exhibits functional alterations in several other cancers [468]. Interestingly, deletion of *DNMT3B* alone leads to only a minor loss of methylation in cell lines; the loss is much more dramatic when both *DNMT3B* and *DNMT1* are disrupted [384]. On the other hand, *TET2* mutations are also frequent in myeloproliferative neoplasms (MPN) [469] and other myeloid malignancies, including AML [470].

To a large extent, methylation patterns in cancer cells recapitulate those of their cells of origin, albeit with a progressive gain of methylation in CGIs and loss outside CGIs [471,472]. In fact, it is possible to define classifiers that accurately predict the tissue of origin on the basis of methylation data [473,474]. However, altered methylation does not only reflect the status of the cell of origin. Ultimately, the loss or gain of methylation at certain loci is contingent on whether these changes provide a selective growth advantage [475].

3.4.5 Detection of DNA methylation

The growing understanding of the role of DNA methylation in health and disease has been enabled by the development of appropriate molecular biology techniques [476]. An early method to assess DNA methylation patterns relied on the observation that certain **restriction enzymes** cannot digest methylated DNA [477,478]. The technique was later refined by the utilization of pairs of enzymes with the same target sequence, but different sensitivity to DNA methylation [479,480]. Building on this principle, the HELP assay (HpaII tiny fragment Enrichment by Ligation-mediated PCR) compares the profiles generated by HpaII and MspI to detect differentially methylated regions [481]. However, restriction enzyme approaches are limited by the impossibility to amplify DNA *in vitro* without losing methylation information and their incomplete coverage of the genome [482].

A key breakthrough came with the adoption of **sodium bisulfite conversion,** a compound that selectively binds and deaminates pyrimidine bases [483]. Cytosine forms an adduct with bisulfite that undergoes deamination to form uridine, whereas 5mC deamination yields thymine [484,485]. Crucially, the latter reaction is two orders of magnitude slower, as the presence of the methyl group inhibits the formation of the bisulfite adduct [486]. This difference in reactivity was leveraged by Frommer and colleagues to identify 5mC residues in conjunction with PCR and sequencing [487]. Under specific conditions, bisulfite treatment converts unmethylated cytosines to uracil, which is then amplified as thymine, whereas 5mC remains nonreactive and is amplified as cytosine.

This basic principle has been successfully exploited to detect methylation in a variety of applications [482], notably in combination with oligonucleotide arrays [488]. Among these, the Illumina HumanMethylation450 BeadChip, which allows the analysis of >450,000 sites [419], has been widely used for methylome characterization in large-scale projects such as the Cancer Genome Atlas (TCGA) [2]. Nevertheless, the gold standard for DNA methylation analysis is **whole-genome bisulfite sequencing (WGBS),** which can evaluate the 28 million CpG sites in the human genome at base resolution [346]. This approach has been employed to accurately chart the methylation landscape across different tissues [343], but it is very inefficient because only 20-30% of the sequencing reads provide relevant information about CpG methylation. A cheaper alternative is Reduced Representation Bisulfite Sequencing (RRBS), which enriches for fragments with a CpG at each end by digesting 5'-CCGG-3' sites with the methylation-insensitive restriction enzyme MspI [489].

Another set of techniques is based on the use of **antibodies that specifically recognize methylated** sequences, a notion originally established in the 1980s [490]. A more recent iteration of this principle is Methylated DNA ImmunoPrecipitation (MeDIP), which captures methylated DNA fragments for subsequent analysis with microarrays or sequencing [491]. Similarly, Methyl-CpG ImmunoPrecipitation (MCIp) exploits the methyl-binding ability of MBD2 proteins in combination with the Fc tail of a human immunoglobulin to identify methylated DNA with high affinity [492]. Both MeDIP and MCIp are affordable and simple

alternatives to WGBS for the screen of large numbers of samples, but are hampered by their low resolution.

Aside from its elevated cost and relative inefficiency, another disadvantage of bisulfite conversion is its induction of DNA damage, which results in loss of material and fragmentation [483]. This limitation has been recently addressed by the use of sequencing technologies capable **of direct detection of DNA methylation**, such as Oxford Technologies' Nanopore [493]. Remarkably, since these technologies produce long reads from single DNA molecules, they can simultaneously assess DNA methylation and variation over the span of kilobases.

The new frontier is the generation of **single cell methylomes**, a challenging effort given the shortcomings of bisulfite conversion. Some authors have tried to circumvent this problem by using restriction enzymes instead [494], but these approaches have their own drawbacks, such as lack of coverage. More frequently, protocols are adapted to minimize the impact of degradation, with changes such as post-bisulfite adaptor ligation [495,496]. Using this strategy, Farlik and others tracked changes in methylation along hematopoietic differentiation and showed that DNA methylation predicts cell type despite low correlation with gene expression [497].

3.5 Regulatory elements involved in transcription

3.5.1 Promoters

Originally identified by Monod and colleagues in 1964, a promoter is a start signal at the beginning of a gene that directs RNA pol II to initiate transcription [91,498]. The minimal stretch of DNA sufficient to direct this process is known as the **core promoter**, defined as a 50-bp region around the TSS that docks the pre-initiation complex [499]. Moreover, the rate of RNA pol II initiation can be modulated by the integration of signals from TFs and co-activators that bind a larger "proximal promoter" region upstream of the TSS [499]. Core promoters can be classified as "focused" if they have a single TSS or as "dispersed" if they contain multiple TSSs in a broad region [92]. Focused initiation preferentially occurs in cell type specific genes under strict regulation, whereas dispersed initiation is associated with constitutively expressed housekeeping genes. Disperse (or broad) promoters, much more abundant, are associated with CGIs and, therefore, are susceptible to regulation via methylation [421,500]. Although focused promoters are a minority in vertebrates, they have been more thoroughly studied given the biological significance of the genes under their control [501].

Core promoters typically contain **one more elements that enable recognition by GTFs** and assembly of the PIC (Figure 11A) [502]. The most well-studied element is the TATA box, located 30 bp upstream of the TSS in focused promoters, but in fact it is only present in 10-15% of gene promoters in mammals [500]. The TATA box is bound by the TATA-box binding protein (TBP), one of the components of the TFIID complex. The most abundant core promoter element is the initiator (Inr) motif, which directly overlaps the TSS [500,503]. The Inr

element is recognized by the TAF1/2 subunits of TFIID and can be found either alone or in combination with a TATA box or other core promoter elements [502]. In the absence of a TATA box, Inr is often accompanied by the downstream promoter element (DPE), positioned +28 to +33 relative to the TSS and also recognized by TFIID [501]. Other promoter elements found in humans include, among others, the downstream core element (DCE), contacted by the TAF1 subunit of TAFIID; the TFIIB recognition elements (BREs) [92]; and the X Core Promoter Elements (XCPE), which directs TBP binding in TATA-less promoters [504].

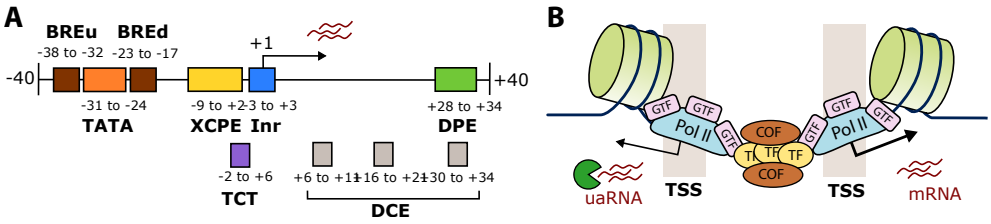


Figure 11. Organization and function of core promoters. (A) Core promoter elements and their sequence motifs (adapted from [501]). (B) Bidirectional transcription from core promoters (adapted from [499]).

Active promoters are characterized by **nucleosome depletion** [114] **and enrichment of histone modifications** in downstream nucleosomes, such as H3K4me3 and H3K27ac [286,505]. These features, together with the presence of a TSS, are currently used for the genome-wide detection of promoter regions. From a functional perspective, it is thought that H3K4me3 may act as a memory mark of recent activity to facilitate future transcriptional events [92]. Bivalent promoters marked by both H3K27me3 and H3K4me3 are inactive, but presumably primed for fast activation [271]. It has been recently shown, however, that H3K4me3 does not confer faster activation, challenging this model [448]. Instead, H3K4me3 may protect H3K27me3-marked regions from irreversible silencing by *de novo* DNA methylation.

Nearly 80% of active promoters, especially broad promoters with CpG islands, exhibit bidirectional transcription from divergently oriented TSSs on opposite strands [500,506,507]. In addition to the gene TSS that produces stable mRNAs, these promoters contain an upstream TSS that generates short upstream antisense RNA (uaRNA), which are quickly degraded by nuclear exosomes (Figure 11B). TFs binding the proximal promoter are located between these two divergent TSSs [499].

3.5.2 Enhancers

Even though core promoters are capable of driving autonomous transcription, they are often weak and require input from distant **enhancers** to reach the gene expression levels required by the cell [508]. Enhancers are DNA sequences of a few hundred bp that contain TF binding sites and **increase the level of transcription from their target promoters** [92,99]. Thus, the activation of the correct subset of enhancers is a critical determinant of cell

identity [273]. Enhancers were discovered in the 1980s through the identification of a 72-bp DNA sequence from the SV40 virus that increased transcription of a reporter gene by ~200-fold, independently of distance and orientation [509,510]. The first cellular enhancer was later found in the immunoglobulin heavy chain (IgH) gene locus, within the intron preceding the constant region [511,512].

These early discoveries revealed some **key hallmarks of enhancers**, namely that they augment gene expression and act independently of distance and orientation to their target genes. Further work showed that they are characterized by a number of epigenetic features, including open chromatin, clustered binding of TFs and enrichment for H3K27ac and H3K4me1 histone modifications, with comparatively low H3K4me3 levels [99]. Bioinformatics predictions based on association with these epigenetic marks, measured by methods such as ChIP-seq ofr ATAC-seq, have identified hundreds of thousands of putative enhancers [513]. Moreover, enhancers are transcribed as eRNAs, often in a bidirectional manner, at levels that correlate with mRNA synthesis by their target genes (Box 3) [514]. The ability to produce bidirectional transcripts has also been exploited to detect active putative enhancers [515,516]. However, it is important to keep in mind that **validating predicted regions as bona fide enhancers requires functional characterization**, proving they can indeed increase transcription from a reporter gene. This task can be accomplished using high-throughput reporter assays, such as CRE-seq [517] or STARR-seq [518]. In CRE-seq, the putative enhancers are inserted upstream of a minimal promoter in barcoded plasmids, whereas in STARR-seq they are inserted in the 3' UTR of the reporter gene, avoiding the need for barcodes. A study using CRE-seq determined that 26% of the ENCODE predictions in K562 have regulatory activity [519]. Among the various epigenetic marks, H3K27ac is the best predictor for active regulatory regions validated by STARR-seq [520], but eRNA levels seem to be more indicative of enhancer activity [521].

Box 3. Functional roles of enhancer expression in gene regulation.

Enhancer-derived RNAs (eRNAs) are generally bidirectional, unspliced and non-polyadenylated [514], although a recent study in single cells concluded that this bidirectionality is an artifact of bulk data [522]. Three main models have been proposed to explain the role of eRNA in gene regulation, reviewed in more depth in [99]. First, both the transcription of enhancers and the resulting eRNAs are non-functional and merely a byproduct of high RNA pol II concentrations. Second, the act of transcription participates in the remodeling of chromatin, by carrying histone transferases or opening up chromatin, even though the resulting eRNAs would be irrelevant. Third, eRNAs themselves have a function, such as the stabilization of enhancer-promoter looping, the binding of TFs or the sequestration of transcriptional repressors.

Enhancer states can be classified as **inactive, primed, poised or active**, each of which is associated with distinct chromatin marks (Figure 12) [273]. Inactive enhancers are located in compact chromatin and thus are inaccessible to transcription factors and cofactors, which results in lack histone modifications. Pioneer factors can bind the DNA wrapped around chromosomes and recruit chromatin remodelers to make the region accessible to other transcription factors and epigenetic modifiers [99,272]. This process, known as enhancer priming, is accompanied by acquisition of H3K4me1 and loss of DNA methylation. Poised enhancers are a category of enhancers associated with lineage specification marked by both H3K4me1 and H3K27me3. An enhancer becomes fully active upon recruitment of GTFs and RNA pol II, leading to initiation of transcription, and HATs (e.g. CBP/p300) that deposit acetylation marks. Acetylated histone tails are bound by BRD4, which activates P-TEFb for RNA poll II CTD phosphorylation, resulting in eRNA elongation. Finally, active enhancers can be decommissioned by a process that involves release of TFs, removal of histone marks, loss of chromatin accessibility and gain of DNA methylation [272].

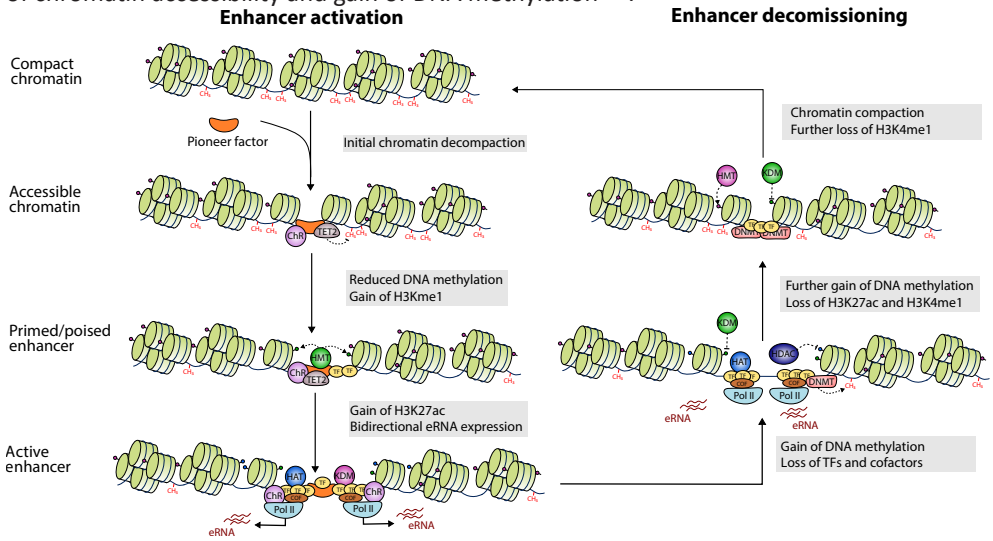


Figure 12. Enhancer activation and decommissioning (adapted from [272] and [99]).

Enhancers and promoters are separated by distances ranging from a few hundred bases to one megabase, a phenomenon made possible by chromatin looping [168,172]. The majority of enhancers are located within 500 kb from their target promoters, which only in a fraction of the cases (between 30% and 60% depending on the study) are the nearest elements [523]. Moreover, a single promoter is often under the regulation of 4-5 enhancers, possibly alternating along differentiation; in turn, enhancers interact with 2 promoters on average [523]. These observations raise questions about the determinants that drive enhancer-promoter specificity, aside from proximity in the linear genome. A crucial requirement in the selection of an enhancer among the repertoire of potential enhancers is its cell-type specific

activation by TFs [273]. Two other important factors are spatial architecture and biochemical compatibility [523]. For enhancers and promoters to interact, chromatin loops must be formed with the aid of specialized architectural proteins – such interactions are typically restricted within TADs. However, even forced contacts between an enhancer and a core promoter are not always sufficient to activate transcription, suggesting they must be compatible as well [92]. Thus, different classes of promoters, possibly depending on their sequence composition, may require specific TF and cofactors that are only present at certain enhancers. Altogether, this cautions against assigning an enhancer to the closest promoter. Instead, other data may be used: a) chromatin interactions derived from 3C technologies, especially Hi-C or promoter-capture Hi-C, and b) correlations between features of promoters and putative enhancers, such as open chromatin or H3K27ac.

Core promoters receive regulatory input from enhancers in the form of TFs and transcriptional cofactors, which **modulate the transcriptional output of the promoter in multiple ways** [92]. Some of these proteins contribute to the assembly and the stabilization of the PIC and the recruitment of RNA pol II, as is the case of the Mediator complex [524] and p300/CBP [525]. However, some core promoters autonomously recruit high levels of RNA pol II and their limiting factor is elongation. In these cases, their cognate enhancers are likely to display high levels of proteins involved in pause-release, such as BRD4 [526] and p300/CBP [525]. Finally, while transcriptional burst size is a fixed property of the core promoter, the frequency of these bursts can be increased by developmental enhancers [527].

The **similarities between enhancers and promoters** have blurred the boundaries between the two classes of regulatory elements. Aside from similar epigenetic marks and bidirectional transcription, enhancers also recruit GTFs [528] and contain core promoter elements like the TATA box [516]. Furthermore, intragenic enhancers can frequently act as alternative tissue-specific promoters [529] and promoters may also have enhancer activity [530]. Altogether, an emergent hypothesis is that enhancers and promoters are in fact the same type of regulatory element, whose primary function depends on the genomic context [506,531]. Nevertheless, differences exist between the two, notably the fact that only promoters produce stable mRNA transcripts, whereas eRNAs from both strands are quickly degraded. This is largely due to a depletion of polyA sites and an enrichment of 5' splice sites downstream of mRNA TSSs [499]. In addition, enhancers initiate less transcription and have less enhancer responsiveness, which may be due to the degeneracy of their sequence-encoded core promoter elements [92].

3.5.3 Super-enhancers

Super-enhancers are **clusters of enhancers characterized by very high levels of transcriptional activators** and chromatin modifications, often involved in the regulation of cell identity genes and oncogenes [532,533]. Contrary to "conventional" enhancers, which have a clear functional definition, super-enhancers were identified based on bioinformatic analysis

of ChIP-seq data by the ROSE algorithm (Figure 13) [532]. In the original publication, the authors first stitched enhancers within 12.5 kb of each other and ranked these clusters, as well as any remaining individual enhancer, by MED1 (part of the Mediator complex) binding levels measured by ChIP-seq. After plotting these values, regions to the right of the inflection point of the curve were considered as super-enhancers. Since then, they have been also defined based on other epigenetic features associated with active chromatin, particularly H3K27ac [534]. Other similar entities, partially but not completely overlapping, have been proposed independently, such as "stretch enhancers" [535] or "locus control regions" [536].

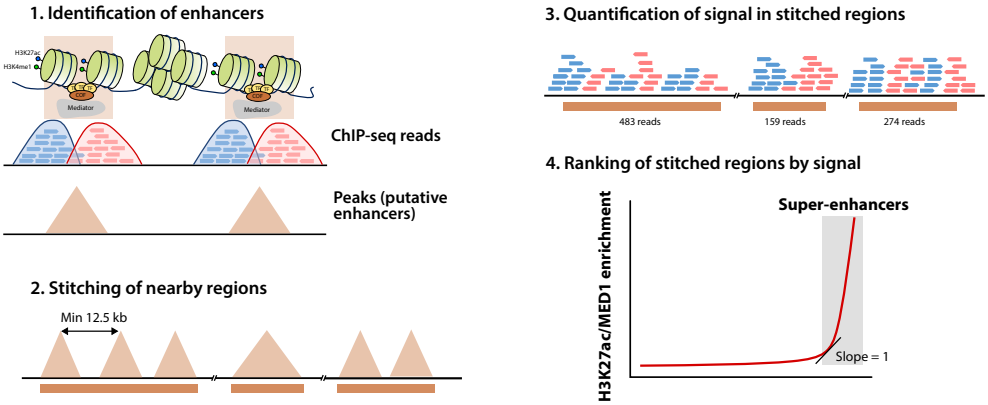


Figure 13. Bioinformatic definition of super-enhancers (adapted from [534]). The identification of enhancers and signal quantification can be conducted with different types of ChIP-seq data, such as MED1 or H3K27ac.

Although the exact number varies across tissues and cell types, an analysis of H3K27ac data from 86 human tissues revealed that **cells have an average of 678 super-enhancers** (median: 657), with an average length of 36345 bp, versus 5154 bp of normal enhancers [537]. Genes under the control of super-enhancers are enriched for lineage-specifying TFs and oncogenes, whose dysregulation may be due to the acquisition of novel super-enhancers in tumor cells [532]. For example, mutation of a non-coding intergenic element results in the formation of an oncogenic super-enhancer in T-ALL, leading to *TAL1* upregulation [538]. Alternatively, genomic rearrangements may bring a super-enhancer into the proximity of an oncogene, as shown for a translocated *GATA2* super-enhancer that upregulates *EVI1* in 3q26 AML [230]. Notably, oncogenic super-enhancers typically contain high levels of BRD4, which makes them susceptible to BET inhibitors such as JQ1 [533].

The majority of super-enhancers (84% in embryonic stem cells) are contained within CTCF-bound cohesin loops that confine their activity to specific target genes, contrary to normal enhancers (48% in the same study) [160]. Disruptions of these boundaries result in dysregulation of nearby genes that can lead to cancer [181]. The **individual components of super-enhancers often interact** with each other through cohesin loops and establish functional interdependence [539]. The individual components may act additively, redundantly

or synergistically [540,541]. Dissection by genetic manipulation revealed that disruption of constituent enhancers enriched in chromatin interactions ("hub enhancers") destabilizes the whole super-enhancer, suggesting a **hierarchical model of organization** [539,541]. A genome-wide study integrating Hi-C and ChIP-seq showed that hierarchical enhancers account for roughly 25% of the total and are particularly associated with genes involved in cell identity [542].

Super-enhancers can also engage in interactions with one another [543] that are largely independent of cohesin [213]. Accordingly, these associations become more frequent following depletion of cohesin, leading to downregulation of genes under the control of super-enhancers. This is consistent with a model in which **super-enhancers form condensates by phase separation** as a result of the high concentration of transcriptional coactivators [544]. These condensates compartmentalize and concentrate the transcriptional machinery, allowing robust expression of target genes, but also interaction between super-enhancers upon fusion of condensates. In Hi-C contact maps, super-enhancers (and some regular enhancers) near a strong loop anchor usually appear as stripes, reflecting a "reeling-in" mechanism whereby one end of the loop remains constant while the other slides in the opposite direction [203]. Moreover, super-enhancers may act as cohesin-loading sites owing to their multiple NIPBL sites [545].

3.5.4 Silencers

Silencers are a class of ***cis*-regulatory elements that reduce transcription** from their target promoters [95]. Like enhancers, silencers function independently of position or orientation with respect to the promoter, their repressive activity is maintained when moved to other locations and can act at relatively long distances [546]. Moreover, they are largely cell-type specific [547]. Since their identification in the 1980s, silencers have remained relatively understudied compared to enhancers, but they play important roles in cell differentiation and disease. For example, a silencer in the first intron of CD4 dynamically regulates the expression of the gene during T cell differentiation [548]. This element is active in double negative (DN) thymocytes and cytotoxic thymocytes, but it becomes inactive in the double positive (DP) stage and in helper lineage cells, thus allowing the expression of CD4.

The nature of silencers is not well understood, but they are generally defined as **open chromatin regions that contain binding sites for transcriptional repressors** such as EVI1 [95]. Some authors have used H3K27me3 to identify putative silencer sites, on the basis of the association of this histone PTM with the binding of transcriptional repressors [549,550]. Huang et al. integrated H3K27me3 ChIP-seq, RNA-seq and Hi-C data to identify putative silencers whose H3K27me3 signal was negatively correlated with interacting genes. Similarly, Ngan and colleagues analysed interactions mediated by PRC2 (which deposits H3K27me3) using ChIA-PET to uncover silencers [551]. However, parallel reporter assays did not reveal a unique association with H3K27me3, suggesting that these regions may constitute only a fraction of the whole catalogue of silencers [547,552]. Other silencers were marked by H3K9me3, which is

often found in heterochromatin regions, and active histone marks such as H3K36me3 and H3K79me2 [547]. Since no single combination of marks accurately discriminates silencers, it has been proposed that they comprise multiple subclasses [95]. Alternatively, the epigenetic signals that identify silencers may not be amongst those that are commonly profiled.

3.6 Transcription factors

Transcription factors (TFs) are proteins that specifically bind to DNA sequences, known as TF binding sites (TFBS), in order to regulate transcription [553]. To achieve this goal, TFs rely on their DNA-binding domains (DBD), which recognize their binding site, and effector domains (ED), which interact with the transcriptional machinery [554,555].

The presence of a DBD constitutes a distinguishing feature of TFs compared to other transcriptional regulators that do not bind DNA directly [556]. Based on the similarity of their DBDs, which are well-conserved structures, the **1639 known or likely human TFs can be classified into 25 families** [553]. Among those, 8 families account 75% the TFs, listed here in decreasing order: C2H2-zinc fingers (ZF), homeodomain, basic helix-loop-helix (bHLH), bZIP, Forkhead, nuclear hormone receptor, HMG/Sox and ETS. Approximately 4% of the TF catalogue do not contain any known DBD, which may indicate they belong to a yet undiscovered DBD group. Some of these families are associated with a particular function; for example, homeodomain TFs are often involved in development [556].

The ED mediates the effect of a TF on gene expression**: activating domains promote transcription, whereas repressing domains have the opposite effect** [554]. This classification determines whether the TF that harbors them is an activator, often found at enhancers or promoters, or a repressor, typically binding to silencers. Some DBD families are predominantly activating (bHLH and homeodomain) while others are repressive (ZF-C2H2) [554]. Bifunctional TFs, whose action is contingent on the cellular context, have also been reported. For example, GLI proteins – which harbour both activating and repressing domains – are proteolytically truncated into repressors unless they are activated by Hedhehog signals [557]. Since EDs are less conserved than DBDs, it is rarely possible to predict their function from their sequence, and they need to be studied functionally instead [554].

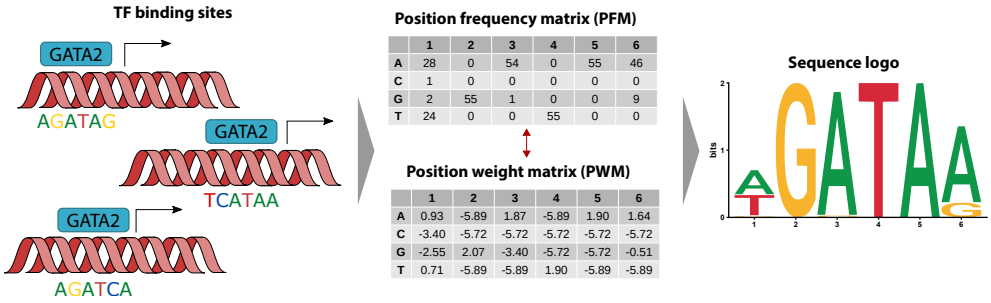
Alternatively, TFs can also be classified depending on their cellular function: while a few are constitutively expressed in all tissues, most are regulated, including lineage-specific TFs (e.g. GATA2) and signal-dependent TFs (e.g. glucocorticoid receptor) [558]. Finally, there are special categories of TFs such as general transcription factors (GTFs), involved in transcription initiation; pioneer factors, capable of binding closed chromatin; or the so-called “master regulators” at the apex of the differentiation hierarchy.

3.6.1 Identification of TFs and their binding sites

The notion of DNA-binding proteins that regulate gene expression can be traced back to the operon-repressor model of Jacob and Monod, according to which repressor molecules

associated with "operator sites" on the DNA to control the synthesis of genes [85]. Shortly afterwards, it was confirmed that the lambda [559] and the lac [560] repressors bound specific sites on the DNA and blocked transcription. In eukaryotes, general transcription factors were identified and isolated in 1980 from HeLa cells [561], the same system in which the first promoter-specific TF, Sp1, was discovered in 1983 [562]. These findings set the foundations for our current understanding of the control of gene expression by TFs binding to *cis*-regulatory elements [563] and kickstarted a series of studies that culminated in the identification of the major TF families by the end of the 1980s [564].

Techniques such as HT-SELEX, ChIP-seq and DNAse-seq have enabled the experimental determination of DNA sequences recognized by TFs, typically summarized as **motifs** [553,556]. Motifs can be represented as a position frequency matrix (PFM) that contains the number of times each nucleotide is observed at every position [565]. Usually, PFMs are converted into position weight matrices (PWMs) that indicate the relative probability of each base with respect to the background in a log scale. For visualization of these patterns, motifs are displayed as "sequence logos" where the size of each base is proportional to its relative occurrence in that position (Figure 14).



	1	2	3	4	5	6
A	28	0	54	0	55	46
C	1	0	0	0	0	0
G	2	55	1	0	0	9
T	24	0	0	55	0	0

	1	2	3	4	5	6
A	0.93	-5.89	1.87	-5.89	1.90	1.64
C	-3.40	-5.72	-5.72	-5.72	-5.72	-5.72
G	-2.55	2.07	-3.40	-5.72	-5.72	-0.51
T	0.71	-5.89	-5.89	1.90	-5.89	-5.89

Figure 14. *De novo* motif identification. Nucleotides present in each position of sequences bound by a TF are summarized in position frequency matrices (PFM), which can be converted to position weight matrices (PWM). The relative weight of each nucleotide in a motif can be represented as a sequence logo.

The identification of new motifs from a set of sequences bound by the same TF is known as *de novo* **motif discovery** and is conducted by specialized computational tools [566], such as Multiple EM for Motif Elicitation (MEME) [567]. Traditionally, these tools were applied to a small number of sequences, which limited their performance when the sites recognized by the same TF had little similarity [566]. The advent of high-throughput technologies overcame this limitation, generating a wealth of data that enabled more sensitive detection of novel motifs. A prominent example is ChIP-seq, which generates thousands of sequences that mostly contain the TFBS of interest. However, this created the need for more efficient algorithms, like MEME-ChIP [568]. On the other hand, the use of ChIP-seq data is constrained by antibody quality, the possible measurement of indirect binding and the biases of sequence

content [553]. Since false positives remain an issue even with more recent approaches, ensemble methods that combine multiple tools can be an attractive strategy [569].

The motifs identified by these algorithms can be compared to previously characterized motifs, which can be retrieved from databases like JASPAR [570] or CIS-BP [571]. Instead of looking for novel motifs, other tools search for existing motifs in one or more input DNA sequences, an exercise known as **motif scanning**. FIMO, one of the most widely used tools in this category, produces a list of possible motifs at each position of the input sequence, ranked by a p-value that estimates the probability the sequence is not random [572]. Finally, **motif enrichment** analyses try to determine which known motifs are overrepresented in a set of input sequences.

3.6.2 Regulation of gene expression by TFs

By virtue of their ability to specifically recognize certain DNA sequences and regulate transcription, TFs have a unique role as decoders of the genome. This process largely takes place at CREs whose activity is determined, among other factors, by the pool of TFs available in a particular cell type [573]. As a result, the same gene may be activated by different enhancers across multiple cell types, each by a different set of TFs. For instance, the expression of *CEBPA* is regulated by as many as 14 putative enhancers, of which the +42 kb is uniquely active in blood and involves binding of TFs like ERG, PU.1, TAL1, RUNX1 and LMO2 [574]. In B-cells, the genomic occupancy of E2A, EBF1 and FOXO1 changes along development [575].

Since motifs are generally short (between 6 and 12 bp long), they are relatively unspecific – there are thousands if not millions of potential TFBSs along the genome. However, most of them are unoccupied, implying that **other mechanisms drive TF specificity**. The impossibility to predict *in vivo* binding based on motif matches has been dubbed the "futility theorem" [565], but understanding of these mechanisms can make this exercise less "futile". Genomic features such as chromatin accessibility, methylation and the shape of the DNA influence TF affinity for a given site [553]. Integration of open chromatin data can, in fact, dramatically improve the ability of PWMs to predict gene expression [576]. High GC content is also associated with increased TF occupancy, possibly in relation to nucleosome occupancy [577]. Moreover, the base pairs flanking the TFBS may influence TF binding by modifying DNA shape, particularly in the case of polyA or polyT tracts [577]. Aside from the DBD, other regions of the TF are involved in determining binding specificity. Intrinsically disordered regions (IDRs) are frequently present in the structure of TFs and can independently direct them to specific promoters [578] or increase TF specificity by interacting with the DBD in the case of p53 [579].

A major factor underlying TF specificity is the need for concerted action between multiple TFs, which avoids undesired transcriptional noise due to spurious recognition of unspecific matches in the genome. Accordingly, **TFBSs appear in dense clusters at CREs** [580], arranged with precise order, orientation and spacing to ensure that TFs can cooperate effectively [93].

In complex organisms, TF cooperation allows a fine control of transcriptional patterns during differentiation and development, as well as cell-type-specific responses to external signals. For example, the SMAD TFs activated by TGFβ interact with OCT4 in embryonic stem cells, MYOD1 in myotubes and PU.1 in B-cells, each of which binds to different locations and activates different genes [581].

Several **modes of TF cooperativity** have been described, depending on their protein structure and the arrangement of their binding sites [582]. Most often, direct protein-protein interactions stabilize TFs occupying adjacent sites on the DNA, as described early on for the lambda phage repressor [583]. The interacting proteins can form dimers (as is the case of bZIPs and bHLHs), trimers or more complex structures [553]. Synergistic effects take place when various TFs interact with the same co-activators, such as p300, increasing the retention time of the TFs at the CRE [93]. Aside from protein-protein interactions, cooperativity can also occur through DNA if one TF alters its shape or dynamics in such a way that favors the binding of the other TF. Moreover, pioneer TFs can create NFRs and indirectly promote the binding of other TFs. Because a single TF is not always sufficient to evict a nucleosome, pioneer TFs often cooperate with other TFs [93,584].

The **mechanism of action** of a TF depends on its effector domain, which typically recruits other proteins to modulate chromatin accessibility, transcriptional initiation and elongation [585]. A classic example of chromatin opening is the recruitment of chromatin remodellers by pioneer TFs such as GATA1, which interacts with the BRG1 ATPase subunit of the SWI/SNF complex [140,586] (see 3.2.3 Mechanisms of nucleosome eviction). Activating domains, often rich in acidic amino acids, can interact with components of the PIC to stimulate its assembly or its activity [555]. In addition, TFs can recruit factors such as P-TEFb to promote elongation, as is the case MYC [587]. On the other hand, some nuclear receptors recruit corepressors that generate silencing chromatin via HDACs, histone demethylases and remodelers [585]. However, bacterial TFs and some human TFs lack an ED and rely on steric hindrance, preventing other proteins from binding that same site [553].

3.7 Epigenetic mechanisms in hematopoiesis

Cell identity along the hematopoietic continuum emerges from the interplay between the different components of the epigenetic landscape, which collectively govern the transcriptional program of cells in response to both intrinsic and extrinsic factors [87]. These epigenetic factors balance the self-renewal and quiescence of HSCs, whereas in differentiating cells they ensure the appropriate expression of cell-type-specific genes. The epigenetic landscape is progressively altered along differentiation, mirroring the observable changes in the phenotype: chromatin becomes accessible or closed, DNA is methylated or demethylated, histones tails are modified, loops between enhancers and promoters are formed or lost. In turn, these changes are instructed by a handful of master regulator TFs that dictate fate choices and maintain cell identity, and whose expression is exquisitely regulated [588].

Transcription factors

Extensive research has gone into identifying TFs essential for hematopoiesis, usually through either gene deletion or forced expression in animal models [8]. High throughput technologies like microarrays and RNA-seq have revealed dense regulatory circuits under the control of these TFs [589–591]. Some of them are involved in the formation and maintenance of HSCs, such as GATA2 [592], RUNX1 [593] or TAL1 [594]. Others play pivotal roles in cell fate specification, by selectively binding the enhancers and promoters of genes associated with a lineage while preventing the expression of genes from alternative lineages. These master regulators are "primary determinants" of cell fate that often function as pioneer factors, reshaping chromatin to facilitate the binding of other TFs (such as EGR or GFI1), which are often "secondary determinants" [595]. Expression of master regulators is often sufficient to direct differentiation into a particular lineage and even force reprogramming of a committed cell into a different lineage [8].

For example, **PU.1 (encoded by *SPI1*) is required for both myeloid and lymphoid development**, but not for the formation of erythrocytes and megakaryocytes [596,597]. Low levels of PU.1 induce B-cell differentiation, whereas high levels promote myelopoiesis at the expense of other lineages [598–600]. Although PU.1 remains expressed in early T-cell precursors up to the DN2 stage, its downregulation is necessary for terminal T-cell maturation [601]. Thus, forced expression of PU.1 in early T-cells can reprogram them into dendritic cells by opposing GATA3, though its effects can be counteracted by NOTCH signaling [602,603]. **Members of the C/EBP family are also critical regulators of myelopoiesis**, with fluctuating expression patterns along this trajectory [604]. In particular, C/EBPA is considered a master regulator of granulopoiesis, as it induces granulocyte formation at the expense of the monocytic pathway in GMPs [605,606]. However, low levels of C/EBPA are necessary for monopoiesis as well as GMP formation [607,608]. **Differentiation into the erythroid and megakaryocytic trajectories is directed by GATA1 and GATA2**, which suppress the myeloid program [609–611]. Both GATA proteins block the action of PU.1 by disrupting its interaction with coactivators, whereas PU.1 reciprocally inhibits the action of GATA1 and GATA2 [612].

Cooperative and antagonistic interactions between both primary and secondary TFs ultimately define lineage commitment, as illustrated by some of the examples above (Figure 15). Thus, the cross-play between PU.1 and GATA1/2 resolves the choice between myeloid and erythroid/megakaryocytic differentiation at the GMP stage, whereas NOTCH1 and GATA3 oppose PU.1 in late T-cell development. Moreover, both C/EBPA [613] and GFI1 [614] inhibit the action of PU.1 to enable granulocyte differentiation at the expense of the monocytic lineage. Similarly, T-bet negatively regulates GATA3 and permits Th1 development instead of Th2 in $CD4^+$ T-cells [615]. Cooperation can take place between TFs that antagonize each other in alternative contexts, as is the case of C/EBPA and PU.1 at the promoter of the GM-CSF receptor (*CSF2R*) in early myelopoiesis [616], or GATA2 and PU.1 in the generation of mast cells [617]. In HSPCs, **a heptad of TFs (TAL1, LYL1, LMO2, ERG, FLI1, GATA2 and RUNX1)**

frequently bind the same genes involved in HSC function and cooperate to regulate their expression [618]. The complexity of these relationships can be more adequately represented as gene regulatory networks, which integrate the knowledge of transcriptional regulation with genome-wide expression data[619,620]. The order in which TFs become expressed [621] and their relative expression levels [622] are also decisive factors in lineage commitment.

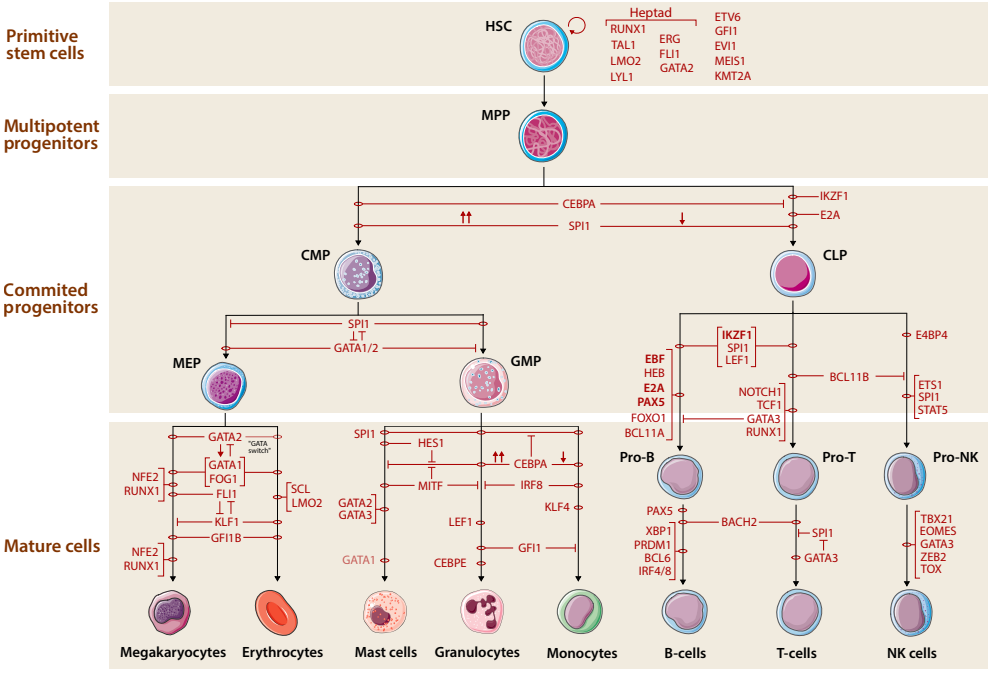


Figure 15. Essential TF for hematopoietic development (adapted from [8]). This diagram presents a simplified, incomplete view of the complex regulatory networks underlying hematopoiesis. Transitions between cell states are indicated with a black arrow. TFs are depicted in red, with activation shown as a circle or an arrowhead, and inhibition as a flathead. A double flathead indicates cross-antagonism between TFs. Vertical arrows indicate high (upwards) or low (downwards) TF expression levels.

Low levels of lineage-specific TFs are coexpressed in HSPCs, suggesting the existence of **multilineage priming** whereby multiple differentiation trajectories are available in early progenitors and they progressively become restricted [8,623]. Although open chromatin and histone modification data partially support this view (see below), new regulatory regions are also created during differentiation. **According to the so-called "instructive" model**, cytokines activate lineage-specific TFs to drive these changes and direct differentiation towards one particular lineage. For example, G-CSF supports neutrophil differentiation of GMPs by increasing the expression of C/EBPA with respect to PU.1, whereas M-CSF promotes macrophage production by inducing PU.1 [622,624]. Likewise, erythropoietin directs erythropoiesis in the HSPC compartment [625]. However, single cell transcriptional profiles have revealed that lineage commitment may start at the HSC stage, with at least myeloid- and lymphoid-biased subpopulations [46,52,53]. This seems to argue in favor of a **"stochastic"**

model in which cytokines only select for clones previously committed to one lineage by random variations in the TF pool. On the other hand, M-CSF can impose a myeloid cell fate on single HSCs, indicating that cytokines are in fact instructive and may possibly override any previously established bias [626]. Aside from modulating the expression of master regulators, cytokines may also induce epigenetic modifiers like KDMs to facilitate the binding of said TFs [627]. Collectively, the data point towards a hybrid of the instructive and stochastic models in which both internal and external factors determine fate choice.

3D genome organization

Hi-C studies have been conducted in various blood cell populations, revealing major spatiotemporal changes in genome conformation [628]. The transition between fetal and adult HSCs is accompanied by spatial reorganization, with increased compartmentalization, TAD boundary strength and genome-wide change of enhancer-promoter loops [629]. A low-input variation of HiC (tagHi-C) was used to characterize spatial changes in rare HSPC subpopulations and myeloid cells, showing that switches in A/B compartments recapitulate the hematopoietic tree [630]. In line with these findings, loss of *Stag2* in murine HSCs, but not of *Stag1*, impairs differentiation and increases self-renewal [631]. Although compensation by *Stag1* preserves TAD boundaries, the lack of *Stag2* disrupts loops that regulate key hematopoietic TFs.

In T-cell differentiation, reorganization of chromatin structure takes place mostly at the transitions from DN2 to DN3 and DN4 to DP, often preceding changes in gene expression and affecting both intra-TAD connectivity and A/B compartments [632]. *BCL11B*, which becomes expressed in the DN2-to-DN3 transition, facilitates the formation and maintenance of T-cell-specific chromatin interactions. A similar role is played by *PAX5* in B-cells, where it orchestrates changes in genome organization independently of transcription [633].

DNA methylation

In hematopoiesis, maturing blood cells exhibit pronounced changes in methylation patterns, with gains and losses in the promoters of specific TFs that frequently mirror changes in their expression [497,634,635]. For instance, *MEIS1*, involved in HSC quiescence, becomes methylated and silenced in mature cells, whereas myeloid-specific genes such as *TAL1* or *MPO* are demethylated in myeloid progenitors. In general, however, TF binding sites are more methylated in HSCs than in more committed downstream progenitors [497]. Lymphoid progenitors and differentiated cells also exhibit higher methylation than their myeloid counterparts, with promoters of myeloid TFs and their binding sites being frequently methylated [635]. Altogether, this suggests that the myeloid program is the default trajectory, and that methylation acts as safety lock against its accidental activation in the lymphoid lineage. The high specificity of methylation patterns enables the identification of predictive signature regions that can be used to reconstruct the hematopoietic tree [497].

The changes in methylation associated with differentiation require the intervention of the DNA methylation machinery. Consistent with the notion that active demethylation allows derepression of lineage-specific regions, targeting *Tet2* in mice via genetic ablation [398] or shRNAs [636] resulted in increased HSC self-renewal, impaired differentiation with a myeloid bias and expansion of the HSPC compartment. One of the key roles of TET2 in hematopoiesis is to bind cell-type-specific enhancers and demethylate them to enable the binding of lineage-specific TFs such as MYC or ITF2 [400]. Interestingly, *Tet1* is dispensable for hematopoiesis, although its deletion in HSCs results in increased self-renewal and a bias towards the lymphoid lineage, but a block in B-cell maturation [637,638]. Along these lines, targeting other TET genes has confirmed that tissue-specific demethylation is essential for B-cell differentiation and function [639].

Experiments in *Dnmt3a*-null mice revealed that DNMT3A enables HSC differentiation by methylating and silencing regions involved in self-renewal and multipotency, like *RUNX1* or *MYCN* [402]. A previous report had concluded that DNMT3A was dispensable for lineage commitment, but the conflicting result may be due to the less efficient deletion of the *Dnmt3a* gene [640]. DNMT3B is also involved in hematopoiesis, but its functions are overlapping with those of DNMT3A, which can compensate for loss of DNMT3B in knock out models [641]. On the other hand, methylation maintenance by DNMT1 is essential for HSC self-renewal and survival, but also for lymphoid differentiation [642,643]. The myeloid skewing seen in models with *Dnmt1* deficiency [642] and *Dnmt3a* haploinsufficiency [644] is in line with the hypothesis that methylation suppresses the default myeloid program.

In summary, *de novo* methylation by DNMT3A/B switches off the stem cell program in HSCs, whereas maintenance methylation by DNMT1 is required for HSC self-renewal and lymphoid commitment. Active demethylation by TET2 allows the expression of lineage-specific genes.

Chromatin accessibility

Chromatin remodelling plays a critical role in hematopoiesis by determining what genomic locations are available to TFs and the transcription machinery. A study combining ATAC-seq and RNA-seq in 16 major blood cell types demonstrated that chromatin accessibility at distal regions, but not at promoters, is a better predictor of cell type than gene expression [645]. This can be attributed to the fact that the expression and promoter accessibility of a given gene can remain invariant along differentiation, yet be controlled by different active enhancers in each cell type. Moreover, changes in chromatin accessibility may precede gene expression, providing discriminatory power earlier in the hierarchy. Subsequent work using single cell data showed that the chromatin landscape in hematopoiesis forms a continuum, largely driven by master lineage regulators that increase their activity in a gradient along differentiation trajectories [646]. Furthermore, progenitor populations were shown to be less homogenous at the epigenetic level than originally thought, with major differences between early and late subgroups.

In general, chromatin is more accessible in HSCs and it becomes increasingly compact as differentiation progresses, in line with a "multilineage priming" model in which CREs of lineage-specific genes are accessible to enable multipotency even before they are expressed [647,648]. Indeed, promoters of genes regulated by master regulators such as GATA1 are often open in HSCs, but not yet expressed [648]. Nevertheless, low level expression of genes linked to multiple lineages, often antagonistic, has also been observed in HSPCs [623]. While a priming model seems to apply to a number of CREs, particularly promoters, cell fate specification also requires *de novo* creation of open chromatin regions at enhancers [647,648].

Chromatin access is primarily facilitated by members of the SWI/SNF family [135]. Loss of *ARID1A,* one of the core subunits of the SWI/SNF complex, leads to global reduction of open chromatin and impaired HSC differentiation, confirming the importance of epigenetic remodelling for hematopoiesis [649]. Similarly, perturbation of other SWI/SNF subunits like *BRG1* [650], *SMARCD2* [651] or *ARID2* (*Baf200)* [652], among others, results in disruptions of hematopoiesis at various stages. For example, *BRG1* and *SMARCD2* are particularly critical for granulopoiesis. The recruitment of SWI/SNF to genes involved in differentiation is mediated by key TFs such as RUNX1, which interacts with the BRG1 and INI1 subunits of the complex [653]. Indeed, depletion of RUNX1 in Jurkat cells results in loss of SWI/SNF binding to the promoters of *IL3* and *CSF2* (GM-CSF), as well as downregulation of these genes.

Histone modifications

The so-called "histone code" has important implications for the regulation of gene expression in hematopoiesis, as it influences chromatin accessibility and the binding of "reader" proteins that recognize specific marks. Thus, profiling of H3K4me1, H3K4me2, H3K4me3 and H3K27ac showed that hematopoietic lineage commitment is accompanied by widespread change in the chromatin landscape [654]. Up to 90% of the enhancers change state, of which 60% are active only in HSCs and their specific lineage, and the rest are established *de novo* during differentiation. This is consistent with findings from ATAC-seq experiments revealing that hematopoiesis follows a hybrid model of increasing restriction of multilineage priming and *de novo* enhancer activation [647,648]. While enhancer decommissioning is a gradual process, *de novo* formation mostly occurs at key transitions, with CMP and GMP accounting for a large fraction of newly formed enhancers in myelopoiesis [654]. Furthermore, the acquisition of different histone marks takes place in a sequential manner, typically starting with H3K4me1/2 at poised enhancers in early progenitors and following with H3K27ac as soon as transcription starts. Several genes involved in HSC differentiation are bivalent regions marked by H3K27me3 and H3K4me3, of which 20% lose H3K27me3 and become active with differentiation and 24% remain bivalent [655]. Bivalent promoters of T-cell master regulators *TBX21* and *GATA3* in $CD4^+$ T-cell may explain the plasticity of these cells [656]. However, the majority of bivalent promoters lose H3K4me3, in keeping with the notion that epigenetic silencing locks out alternative lineages [655].

The PRC2 complex, whose catalytic subunit is one of the KMTs EZH1 or EZH2, catalyzes the **methylation of H3K27**. In fetal hematopoiesis, this process is exclusively dependent on EZH2, since EZH1 is not expressed [657]. In adult HSCs, EZH2 is dispensable for self-renewal, but it is involved in lymphoid differentiation, particularly of B cells [657,658]. Deletion of *Ezh2*, however, enhances NK cell commitment and cell function in a cell-intrinsic manner [659]. On the contrary, EZH1 is essential for HSC self-renewal and quiescence, by selectively repressing *CDKN2A*, but also for appropriate lymphoid differentiation and B-cell development [660]. The strong dependency of lymphoid differentiation on H3K27me3-mediated silencing is in line with the notion that the myeloid trajectory is the default. Deletion of the core subunit *Eed* (shared by EZH1- and EZH2-containing complexes) has an even stronger impact on HSC differentiation and survival, suggesting that each type of complex has specific targets [661]. Interestingly, mutations in any of the PRC2 subunits result in enhanced HSC proliferation [662], yet overexpression of EZH2 also increases HSC self-renewal and prevents exhaustion [663] and complete loss of EED leads to HSC exhaustion [661]. Therefore, normal HSC function requires fine regulation of H3K27 methylation levels by PRC2.

The KDMs **UTX and JMJD3 de-methylate H3K27** at early stages of differentiation to enable the binding of lineage-determining TFs to target regions [627]. Accordingly, pharmacological inhibition of these enzymes prevents TF binding and blocks cytokine-induced differentiation. *Utx*-deficient mice models exhibit increased HSC self-renewal with a myeloid bias stemming from inhibited expression of erythroid TFs [664]. On the other hand, LSD1 removes H3K4me1/2 at promoters and enhancers of stemness genes, suppressing their expression and thereby permitting HSC differentiation [665]. Since LSD1 mediates transcriptional repression in various contexts, it is also critical for terminal differentiation as well as HSC self-renewal [665,666].

Histone acetylation is catalyzed by HATs, among which **the CBP/p300 family** is indispensable for definitive hematopoiesis, though not for HSC formation in early development [667–669]. The interactome of CBP/p300 includes more than 400 proteins, among others master regulators like C/EBPA, GATA2 or KLF4 [670]. Despite the high degree of homology between CBP and p300, their functions are not identical [671]. These differences may be related to their different specificities for the same histone H3 and H4 residues [672]. Specifically, full complement of CBP is required for HSC self-renewal and maintenance of the HSC pool, with *Cbp* +/- mice exhibiting hematopoietic failure [667,668]. This effect is cell-autonomous, as deletion of *Cbp* in the niche does not alter HSC repopulation ability [673]. Moreover, conditional knockout of *Cbp* in murine HSCs results in bias towards the myeloid lineage, indicating a role for CBP in differentiation [673]. On the contrary, p300 in HSCs is dispensable for both self-renewal and differentiation, whereas loss of p300 in the niche compromises differentiation [667,668]. However, CBP and p300 may compensate for each other in T-cell [674] and B-cell [675] development, which is only blocked by the combined deletion of both p300 and CBP.

The action of HATs is counterbalanced by HDACs, which repress transcription by deacetylating lysine residues. In hematopoiesis, they frequently form complexes with TFs and cofactors that regulate every stage of differentiation, extensively reviewed in [676]. HDAC1 and HDAC2 are essential for HSC maintenance and differentiation, though they can functionally compensate each other to varying degrees [677]. While they have largely overlapping roles in HSC homeostasis and myeloid differentiation, HDAC1 is more dominant in erythropoiesis and thymocyte development [677]. The levels of HDAC1 in MEPs are upregulated by GATA1, whereas they are downregulated in myelopoiesis by C/EBP proteins [678]. Similarly, HSPCs depend on HDAC3 for proliferation as well as differentiation into the lymphoid and erythroid lineages [679]. SIRT3 controls HSC homeostasis by deacetylating H3K56 at WNT genes and inhibiting their expression, thus preventing proliferation [680].

It should be noted that the functions of HDACs and HATs do not necessarily involve histones, as they also deacetylate other proteins. For example, HDAC8 ensures LT-HSC maintenance by inactivating p53 [681], whereas SIRT1 protects HSCs from ageing by promoting the nuclear localization of FOXO3 [682].

4. ACUTE MYELOID LEUKEMIA

Defects in the regulation of self-renewal and differentiation may lead to clonal expansion of immature precursors with impairment of healthy blood production, the pathological hallmarks of acute myeloid leukemia (AML) [683]. Transformation may take place in HSCs or in committed progenitors that acquire aberrant stem cell characteristics [684–686]. The accumulation of immature cells known as myeloblasts is accompanied by loss of granulocytes, erythrocytes and thrombocytes, resulting in severe infections, anemia and bleeding, respectively [683]. For the diagnosis of AML, a minimum of 20% blasts in the bone marrow or blood is required, except when the aberrations t(15;17), t(8;21), inv(16) or t(16;16) are present [687].

An early attempt to distinguish between subclasses of AML was the French-American-British (FAB) classification system established in 1976, which defined groups ranging from M0 to M7 based on their morphological characteristics [688,689]. Although it remained in use for decades, the tremendous progress in understanding of AML from a genetic and biochemical perspective eventually rendered it obsolete. Some FAB subgroups overlap with cytogenetic aberrations, as is the case of M4 with inv(16) and M3 with t(15;17), but overall there is a large disconnect [690]. Attempting to integrate all the available information, the World Health Organization (WHO) developed a new classification system in 2001 [691], last updated in 2022 [692]. The WHO classification identifies AML subtypes on the basis of recurrent abnormalities, such as t(8;21), inv(16) or biallelic *CEBPA* mutations. The same approach is followed by the recent International Consensus Classification (ICC) [693]. In parallel with these classifications, the European LeukemiaNet (ELN) consortium publishes recommendations that stratify risk as “favorable”, “intermediate” and “adverse” on the basis of genetic alterations [687].

Besides, AML can also be classified on the basis of its ontogeny [694]. **Primary or *de novo* AML** is diagnosed in patients without a history of hematologic diseases or treatment with chemotherapy or radiation. In contrast, **secondary AML (sAML)** evolves from a previously diagnosed hematologic disorder, such as myelodysplastic syndromes (MDS), **whereas therapy-related AML (t-AML)** results from exposure to leukemogenic therapies. Primary AML accounts for 74% of all diagnosed cases [695,696], though a fraction of those could be sAML with unrecognized antecedent MDS [694]. Almost 20% of the cases are classified as sAML, of which ~60% derive from MDS, ~25% from myeloproliferative neoplasms (MPNs) and 10% from chronic myelomonocytic leukemia (CMML) [695,696]. The remaining 6-7% of cases are tAML, frequently derived from breast cancer, uterine cancer and other hematological malignancies. Each of these classes can be associated with unique mutational patterns and clinical variables, with sAML being more common in the elderly and associated with inferior survival [694].

AML is the second most common type of leukemia yet the most lethal, accounting for roughly 30% of the newly reported leukemia cases but 40% of the deaths [697–699]. This contrast illustrates the poor prognosis of this disease, which has an estimated 5-year survival of 30% [700].

AML represents 1.1% of the new cancer cases in the US, with 0.5% of the population at risk of developing AML in their lifetime [700]. The incidence of AML exhibits a J shape, slightly decreasing after childhood but quickly increasing after early adulthood, especially in the elderly [698]. Thus, with a median age at diagnosis of 68 years, AML primarily affects older adults [700]. This is a consequence of the progressive accumulation of mutations with age in the HSCs of healthy individuals [701]. The incidence of AML has substantially increased in the last decades, partially due to the ageing of the population, but also due to improved diagnostic techniques and a growing number of therapy-related cases previously treated for another malignancy with chemotherapeutic agents [698].

Even though AML was deemed incurable for years, between 35% and 45% of patients below 60 years of age and 10-15% of those older than 60 currently achieve long term survival [683]. The standard of care is **induction therapy with a combination of cytarabine and an anthracycline**, typically in the form of a 7+3 regime, i.e. 7 days of cytarabine followed by 3 days of daunorubicin [687]. Upon remission, patients receive **consolidation therapy**, which is often intensive chemotherapy in favorable risk groups and allogeneic HSC transplantation in intermediate or adverse risk groups [687]. The 7+3 regime achieves complete remission in 60-80% of patients below 60 and 40-60% in the elderly, provided they are eligible for therapy [687]. In recent years, an arsenal of novel treatments has become available, including hypomethylating agents, immunotherapies and autologous CAR T cells [702,703]. The availability of **targeted therapies together with comprehensive mutational profiling** has opened the door to precision medicine, which promises to improve remission rates and overcome refractory cases. Some notable examples are FLT3 inhibitors for patients with FLT3-ITD mutations and IDH1/2 inhibitors for IDH1/2-mutated AML [702]. Given the prominent role of epigenetic dysregulation in AML, therapies aimed at epigenetic modifiers are of particular interest, but so far only methylation-related treatments have been successful [702].

4.1 Clonal evolution in leukemogenesis

The development of leukemia, known as leukemogenesis, is a stepwise process of selection of cells with characteristics that confer them a fitness advantage [704]. The currently accepted model of clonal evolution in cancer was postulated by Peter Nowell in 1976 [705] and applied to leukemia by McCulloch and colleagues in 1977 [706]. The leukemic cell population constitutes a **clone** that derives from a single **cell of origin**, and the Darwinian selection that results in tumor progression is thus known as **clonal evolution** [705]. The initiating event of this process is a genetic alteration, such as a point mutation or a chromosomal rearrangement, which favors the expansion of a cell over its normal counterparts. This is generally a preleukemic lesion, as it is insufficient to induce full-blown leukemia on its own. Subsequent alterations in the descendants of the founding clone give rise to subpopulations (or **subclones**) that eventually become dominant if favored by natural selection.

The progression from preleukemic states to clonal hematological disorders such as MDS or CML, eventually culminating in full-blown leukemia, is a paradigmatic example of clonal evolution (Figure 16). Although originally proposed in 1977, it was not until recently that high throughput sequencing enabled the reconstruction of this process. Transformation from MDS to AML involves the persistence of a founding clone, although it can be outcompeted by descending subclones [707]. As expected, genetic lesions are accumulated over time, resulting in subclones that carry increasing numbers of mutations, some of which confer growth advantage over the others. Aside from disease initiation, Nowell predicted that clonal evolution plays a critical role in resistance to therapy and relapse, as it allows cancer cells to adapt to the selective pressure of the treatment and expand anew [705]. Indeed, WGS of *de novo* AML confirmed that relapse is associated with acquisition of novel mutations by either the founding clone or a surviving subclone, followed by clonal expansion [708].

Somatic mutations are acquired throughout life as a consequence of cell-intrinsic mechanisms, such as replication errors or cytosine deamination, and exposure to various mutagens [704]. Only a small fraction of those, known as **driver mutations,** confer fitness advantage and are positively selected. The vast majority of variants found in cancer are **passenger mutations** that are not influenced by natural selection and are instead fixated due to random genetic drift, in line with the neutral theory of molecular evolution formulated by Motoo Kimura in 1968 [709]. This is due to the fact that these variants are either located in non-functional "junk" DNA [710,711] or do not alter the aminoacid sequence (synonymous mutations). In Nowell's original model, most coding variants were expected to be removed by purifying selection due to either metabolic disadvantage or destruction by the immune system. However, while this mechanism is pervasive in species evolution, it is largely absent in cancer cells [712,713]. Most non-synonymous mutations are simply tolerated by cancer cells, which can be explained by the presence of two or more gene copies, the dispensability of most genes for a given somatic lineage or the hitchhiking with driver mutations (i.e., their negative effect is offset by the co-occurrence with a positively selected mutation).

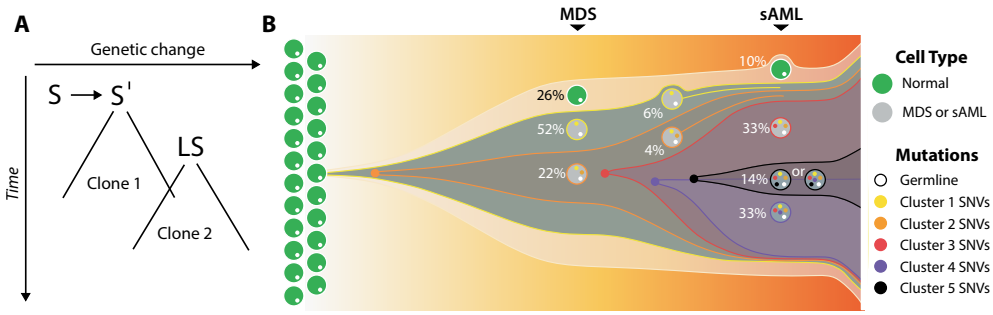


Figure 16. Clonal evolution in leukemogenesis. (A) Original model proposed by McCulloch and others [706]. A stem cell (S) may acquire genetic mutations and become a preleukemic clone (S'), which over time can give rise to a leukemic stem cell (LS). (B) Example of clonal evolution from MDS to AML, characterized using whole genome sequencing [707].

4.2 The leukemic stem cell (LSC) concept and the AML cell of origin

Although AML is characterized by the accumulation of large numbers of blasts, these cells have limited proliferative capacity, suggesting the existence of a small pool of **LSCs endowed with self-renewal and differentiation** potential [706,714]. In 1967, Philip Fialkow and colleagues determined the clonal origin of CML based on the X inactivation of the G6PD locus [715]. Other early studies identified leukemic chromosomal aberrations in multiple hematopoietic lineages, including granulocytes [716,717] and erythrocytes [718], indicating they derived from a pluripotent cell with the ability to differentiate. Subsequent work by McCulloch et al. identified a fraction of the leukemic population with proliferative activity, supporting the notion that the blast population is maintained by self-renewing LSCs [719,720]. Towards the 1980s, the concept of LSCs with the capacity to proliferate and differentiate enoyed widespread acceptance, supported by multiple *in vitro* colony formation experiments [721]. Nevertheless, the AML colony-forming units (AML-CFU) detected by these assays had limited proliferation and replating potential.

It was not until 1994 when a landmark study by John Dick and colleagues characterized *bona fide* LSCs, capable of fully establishing human leukemia in immunodeficient mice [722]. These leukemia-initiating cells were more rare and primitive than AML-CFUs and exhibited a CD34+CD38- immunophenotype, similar to normal HSCs. Along these lines, other groups detected cytogenetic aberrations in CD34+ subpopulations [723]. A follow-up publication by the Dick group further showed that LSCs derived from different patients shared the same CD34+CD38- surface markers [724]. They demonstrated that these cells proliferate and differentiate in primary recipients, and exhibit self-renewal potential in secondary transplants. These data were the basis to propose a **hierarchical model for malignant hematopoiesis in which LSCs generate more restricted progenitors** and large numbers of myeloblasts in a clonal manner. In lien with this notion, Goardon and colleagues showed that immature $CD34^{+}CD38^{-}$ LSCs resembling LMPPs give rise to a population of GMP-like LSCs [686].

Numerous studies have extensively characterized LSCs in different AML subtypes [725]. Although more than 75% of AML cases harbor CD34+ LSCs, others, AML subtypes with mutations in either *NPM1* or *TET2* generally lack CD34 expression and their LSCs are found in a CD34- compartment [725]. A number of other surface markers have been detected, further establishing the phenotypical heterogeneity of LSCs, whose identity may depend on the genetic and epigenetic background of each leukemia [726].

The elusive cell of origin

The LSC concept is intimately linked to the cell of origin of AML. On the basis that LSCs exhibit both self-renewal and differentiation potential, McCulloch reasoned **that transformation takes place in pluripotent HSCs**, which expand following the acquisition of genetic defects [706]. Strong evidence for this hypothesis was the identification of leukemia-initiating cells with

HSC surface markers [722], as well as the observation that CD34+CD38- populations in human AML patients carried cytogenetic aberrations [723]. An alternative model posited that **many different cells in the HSPC compartment are susceptible to transformation**, which would entail the acquisition of aberrant self-renewal potential [721]. This was initially supported by the inter-patient heterogeneity of differentiation trajectories in AML, with only a fraction of the cases retaining erythropoietic capability [727,728]. Moreover, HSCs defined by CD34+CD38- isolated in AML with t(15;17) did not harbor the PML-RARA fusion gene produced by this translocation [729].

The controversy continued well into the twentieth century and remains unsettled. The LMPP-like and GMP-like LSCs identified by Goardon et al. in primary AML may indicate that the cell of origin is a committed progenitor that acquired self-renewal potential, rather than an HSC [686]. Nevertheless, although it is tempting to decipher the cell of origin on the basis of LSC surface markers, those could be aberrantly expressed upon transformation. Thus, functional assays have been conducted to determine the transformation potential of different cell types (Figure 17). Experiments in murine models revealed that **both HSCs and committed progenitors can be transformed into LSCs by the introduction of oncogenic fusion** proteins such as MLL-AF9 [685,730] or MOZ-TIF2 [731]. On the contrary, transduction of BCR-ABL [731] or overexpression of HOXA9 [732] only initiated leukemia in HSCs. Interestingly, MLL-AF9 induces leukemia more rapidly and efficiently in HSCs than in GMPs, implying that the cell of origin may influence leukemia phenotype and clinical outcome in patients [733].

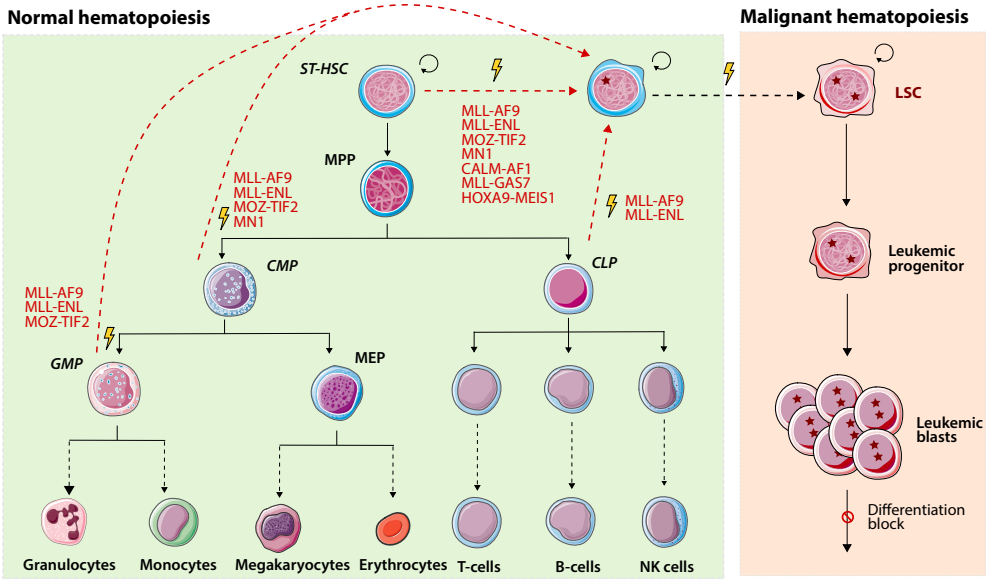


Figure 17. Possible cells of origin in leukemia identified by transduction experiments in mice (adapted from [734]). Cell types that could be transformed by oncogene transduction are highlighted in italics. The different oncogenes that successfully led to leukemia are indicated in red.

On the other hand, several studies have detected **AML-associated mutations in preleukemic HSCs**, suggesting they are the cell of origin that undergoes clonal evolution upon the acquisition of additional mutations. Early evidence came from detecting the expression of RUNX1-RUNX1T1 fusion oncoprotein in HSCs of patients in remission [735]. Screens in residual HSCs from AML patients detected mutations in multiple genes such as *NPM1*, *TET2*, *IDH2* and *DNMT3A*, which were shown to precede later mutations by reconstruction of clonal evolution with single cell data [736,737]. Similarly, preleukemic HSCs with *DNMT3A* mutations and increased repopulation potential were found in the blood of AML patients both at diagnosis and remission [738]. The presence of dominant HSCs long before the onset of leukemia has been termed **clonal hematopoiesis of indeterminate potential (CHIP)** and is associated with increased risk of hematologic cancers and cardiovascular disease [739–741]. The acquisition of mutations in certain genes, most commonly *DNMT3A*, *TET2* and *ASXL1*, promotes clonal expansion of the affected cells to the detriment of the rest of the HSCs. However, only a fraction of these are truly preleukemic cases that go on to develop AML, which can be preemptively identified by unique mutation distributions and greater VAFs [742].

Altogether, the emerging consensus is a **hybrid model whereby AML can originate either in HSCs or in more committed progenitors** [726]. The realization that AML is often preceded by a long preleukemic phase characterized by CHIP strongly suggests that the initial clonal expansion involves HSCs, since short-lived progenitors would be lost in that span of time [743]. Nevertheless, it is possible that the acquisition of additional mutations and resulting transformation take place in a progenitor, which would then acquire stem-like properties. In line with this possibility, the BCR-ABL fusion gene expands the HSC compartment of CML patients in chronic phase, but additional events such as beta catenin activation may transform downstream GMPs and confer them self-renewal potential [744].

4.3 Recurrent mutations in AML

AML is one of the malignancies with the lowest frequency of mutations, only a few of which are drivers [745]. When only coding regions are considered, **AML patients carry an average of 10 to 13 single nucleotide variants (SNVs) and small insertions or deletions (indels)**, of which merely 5 are recurrently identified [2]. AML genomes harbor approximately 2 copy number alterations (CNA) on average [746], and less than 1 fusion gene [2]. Although Nowell hypothesized that the acquisition of mutations in tumor progression is accelerated by genomic instability, this is not the case in AML. Instead, most detected mutations are acquired randomly in the founding clone before cancer initiation in an age-depending manner [701]. Accordingly, HSPCs derived from healthy donors carry a similar number of mutations as their malignant counterparts in AML patients of the same age [701]. Age-associated mutations include spontaneous deamination of 5mC to thymine, indels introduced during the repair of double-strand breaks, polymerase errors and large structural variations [747].

In the early 2000s, Gary Gilliland proposed a **"two-hit model" whereby the development of AML requires co-occurring mutations in two different classes of genes** [748]. Class I mutations stimulate survival and/or proliferation and typically affect signaling genes (e.g. *FLT3*, *KRAS*), whereas class II mutations block hematopoietic differentiation by disrupting TFs (e.g. *RUNX1*). This model was based on the observations that a) progression from CML to AML involved the acquisition of class II fusion proteins like RUNX1-EVI1, b) single fusion oncoproteins were not sufficient to induce leukemia in transformation assays or murine models, c) mutations in the FLT3 tyrosine kinase were found in ~30% of AMLs. In support of this model, the combination of PML-RARA and FLT3-ITD induced leukemia in all transplanted mice, contrary to PML-RARA alone [749]. Moreover, WGS of 24 AML cases revealed that the founding clone frequently requires only two or three cooperating mutations [701]. Individually, mutations in either of these classes can produce other clonal hematological disorders, but not AML. For example, the majority of CML patients carry class I *BCR-ABL* fusions [750–752], which also cause a CML-like disease in mice [753]. Mutations in another signaling gene, *JAK2*, are a frequent cause of MPN [754,755] and similarly recapitulate this phenotype in mice [756]. On the other hand, class II mutations are often associated with MDS, as confirmed by murine models that developed MDS when transplanted with bone marrow cells harboring *RUNX1* mutants [757].

Although this hypothesis provides a useful conceptual framework, it has been challenged by the discovery of mutations in unrelated genes as a part of large-scale sequencing efforts, which depict a more complex landscape [2,758]. Considering their biological function and role in AML pathogenesis, these recurrent mutations can be classified in several categories in addition to the two originally proposed (Table 2, Figure 18). Notably, only ~60% of AMLs carry signaling mutations and ~40% have genetic alterations involving TFs, often as fusion proteins [2]. However, mutations in other genes may achieve equivalent effects by alternative pathways; for example, differentiation arrest can also be a consequence of epigenetic dysregulation. On the other hand, mutations rarely fit neatly in any particular class, since their downstream effects may extend beyond a single pathway. For instance, FLT3-ITD does not only stimulate proliferation, but also represses differentiation via downregulation of *CEBPA* [759]. Alternatively, genetic alterations can be classified on the basis of their association with different AML ontogenies: mutations in splicing genes are specific for sAML, whereas *NPM1* is exclusive of *de novo* AML [694].

Table 2. Functional categories of recurrent mutations in AML. Adapted from [1] with data from [2] and [760]

Functional group	Most commonly mutated genes	Overall frequency
Signaling pathways	*FLT3*, *KIT*, *PTPN11*, *KRAS*, *NRAS*, *BCR-ABL*, *CBL*	52%
DNA methylation	*DNMT3A*, *TET2*, *IDH1*, *IDH2*	43%
Chromatin modifiers	*KMT2A* fusions, *ASXL1*, *EZH2*, *KDM6A*, *BCOR*	30%
Transcription factors	*CEBPA*, *RUNX1*, *PML–RARA*, *MYH11–CBFB*, *RUNX1–RUNX1T1*, *EVI1* rearrangements	33%
Genome organization	*STAG2*, *RAD21*, *SMC1*, *SMC3A*, *PDS5B*	13%
Spliceosome complex	*SRSF2*, *U2AF1*, *SF3B1*, *ZRSR2*	18%
Tumor suppression	*TP53*, *WT1*, *PHF6*	16%
Nucleophosmin*	*NPM1*	23%

* *NPM1* has multiple functions including ribosomal biogenesis, centrosome duplication, DNA damage response, histone chaperoning and transcriptional regulation, among others [761]. Therefore, and given its prominence in AML pathogenesis, it was assigned to a category of its own.

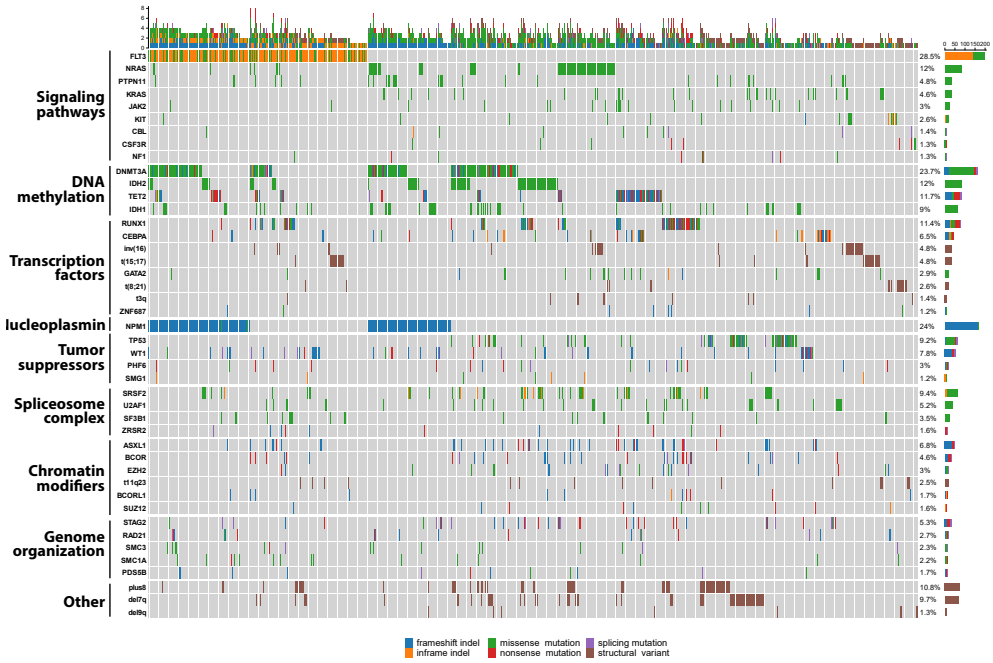


Figure 18. Oncoprint showing the distribution of recurrent mutations in AML patients (more than 1%). Genes are grouped by functional categories and ranked by their mutational frequency. Each column corresponds to a patient, with the type of mutation indicated by a color code described in the legend below. Figure based on data from [2] and [760] (n=732).

The **order of acquisition of these mutations** has been subject of intensive research. By measuring the variant allele frequency (VAF) of each variant, it is possible to infer clonal relationship and thus establish a relative timeline for their acquisition in a patient [701]. Briefly, mutations with the highest VAF (0.5 for a single allele mutations with ploidy 2) were acquired in the founding clone, whereas later subclonal events have smaller VAFs proportional to the size of that subclone. Such studies identified mutations in the epigenetic modifiers ***DNMT3A*, *ASXL1* and *TET2* as some of the earliest events** in AML [736,758], in line with the observation that they are frequently present in pre-malignant HSCs of individuals with CHIP. Mutations in IDH1/2 were also early events, whereas mutations in NPM1, TFs, splicing factors and chromatin were intermediate. Mutations in **signaling genes were generally late subclonal events**, yet also amongst the most common type of mutation. However, in sAML progressing from CML, the acquisition of BCR-ABL is typically the earliest event, indicating that this pattern is not an absolute rule. Interestingly, different orders of acquisition may lead to different disease phenotypes and clinical outcomes, as shown for *JAK2* and *TET2* mutations in MPN [762].

Although coding mutations are too few in AML to elucidate clonal architecture in detail, this obstacle can be surmounted by the coverage of non-coding regions in WGS [701]. **The patterns of co-occurrence between mutations are another source of valuable information**, as they suggests possible mechanisms whereby different pathways collaborate in leukemogenesis and further our understanding of the order of acquisition [758]. For example, the presence of *TP53* mutations in AML with complex karyotype indicates that the absence of this tumor suppresor allows the subsequent accumulation of chromosomal abnormalities that would otherwise trigger cell death. Alternatively, mutually exclusive mutations often involve genes with redundant functions, such as IDH1, IDH2 and TET2, all of which are involved in demethylation. **More detailed information about the clonal substructure can be gleaned from single cell sequencing data** [763–765]. Besides validating conclusions inferred from bulk-sequencing data, such as the subclonality of *NPM1* mutations [764], single cell genotyping can unmask hidden clonal relationships, like the mutual exclusivity between *TP53* and *PPM1D* mutations [763], and reveal instances of both convergent and parallel evolution [765].

4.4 Epigenetic dysregulation in acute myeloid leukemia

Hematopoiesis is a tightly controlled process that involves numerous transcription factors and epigenetic modifiers, which jointly instruct cell-specific transcriptional programs along differentiation. As outlined in previous sections, disturbances in any of these critical factors may severely compromise cell identity and lead to aberrant behavior. In AML, almost 75% of the patients carry recurrent mutations in genes related to epigenetic processes, including DNA methylation, chromatin modifiers, 3D organization and TFs (Table 2). Therefore, it is apparent that epigenetic dysregulation lies at the center of AML pathogenesis, possibly driving the acquisition of cancer hallmarks in HSCs and other progenitors [766].

4.4.1 The effect of mutations in epigenetic regulators

Transcription factors

The association between recurrent **mutations in hematopoietic TFs and dysregulated transcriptional programs** was demonstrated by gene expression profiling [767]. Clustering was largely driven by chromosomal aberrations involving TFs, such as t(8;21) [*RUNX1-RUNX1T1*] or inv(16) [*MYH11-CBFB*], and mutations in the myeloid TF *CEBPA*. Subsequent analyses in larger cohorts further established a subgroup with double *CEBPA* mutations (*CEBPA* DM) as a separate entity with a unique expression profile and favorable disease outcome [768,769]. Mutations in *NPM1*, which is also involved in transcriptional regulation, defined another subgroup [770]. More recent studies detected the same patterns using RNA-seq [760].

Additional research has attempted to elucidate how these mutations alter the transcriptional program of the cell and induce leukemogenesis. The *CEBPA* gene encodes for a full-length p42 (42 kDa) isoform and a shorter p30 isoform, which lacks a fraction of the C-terminal region containing a transactivation domain [771]. Mutations in the N-terminal region *CEBPA* introduce a premature stop codon that leads to the loss of the p42 isoform, required for control differentiation and proliferation, whereas C-terminal mutations affect the bZIP domain, involved in DNA binding and dimer formation [772]. As a hematopoietic TF, C/EBPA binds promoters and enhancers of genes involved in myeloid differentiation and HSC function [773]. The deletion of *CEBPA* results in epigenetic changes at bivalent promoters of genes essential for HSC function [773].

The RUNX1-RUNX1T1 fusion protein generated by t(8;21), also known as RUNX1-ETO, lacks the transactivation domain of RUNX1 and instead recruits HDACs and co-repressors, acting thus mainly as a repressor [774–776]. However, it can also function as a transcriptional activator of self-renewal genes via its NHR1 domain, which is dependent on acetylation by p300 [777]. Up to 60% of the sites bound by RUNX1-ETO are normally occupied by RUNX1, which is displaced upon competition by the oncoprotein [778,779]. Thus, depletion of RUNX1-ETO in the Kasumi-1 cell line restores binding of RUNX1, confirming mutual exclusivity. These experiments also suggest that RUNX1-ETO induces widespread changes in histone acetylation and gene expression, driving self-renewal while blocking differentiation [778]. To a large extent, the reorganization of the transcriptional network that causes this differentiation arrest is driven by the loss of *CEBPA* [779], which is directly repressed by RUNX1-ETO upon binding of the latter to the +42 kb *CEBPA* enhancer [780,781]. Indeed, differentiation block can be overcome by overexpression of *CEBPA* [782].

DNA methylation

Aberrant DNA methylation is a common event in cancer, with **frequent hypermethylation of CGIs accompanied by global decrease of 5mC in the genome** [439]. Early attempts at methylation profiling revealed hypermethylation of several loci in subsets of AML, including

the promoter of the tumor suppressor *CDKN2B* [783,784]. The analysis of 344 AML patients with the HELP assay identified that a **subset of genes are consistently methylated in AML**, possibly pointing to a common involvement in transformation [3]. This study also revealed **16 clusters with distinct methylation patterns**, most of which were enriched for recurrent mutations such as t(8;21) or inv(16), except for 5.

Two of those clusters were later shown to be defined by the presence of mutations in *IDH1* and *IDH2*, both of which inhibit the hydroxylation of 5mC by TET2, leading to global hypermethylation, particularly at promoters [636]. Mutations in *IDH1/2* were shown to be largely exclusive with loss-of-function *TET2* mutations and both groups shared similar methylation signatures, suggesting functional redundancy between *TET2* and *IDH1/2* [469,470]. Likewise, it was later shown that *WT1* mutations are also mutually exclusive with lesions in *TET2* and *IDH1/2*, and also result in decreased TET2 function [785]. Mutated TET2 proteins have less catalytic activity and lead to diminished 5-hmC levels [786], but their effect on global methylation is less clear. Early reports showed both hypomethylation [786] and hypermethylation in *TET2*-mutated patients [636], but they may have been limited by the use of microarray technologies. Bisulfite sequencing revealed that loss of *TET2* induces a hypermethylation phenotype outside CGIs [787], particularly at enhancers [788]. However, the effects of *TET2* mutation are modest overall, in line with studies showing that loss of TET2 in hematopoiesis is compensated by other TET enzymes [789].

The discovery of highly recurrent mutations in *DNMT3A* further highlighted the role of methylation in AML [464], as well as in preleukemic clonal expansion [738]. The most common *DNMT3A* mutation (R880H) results in dominant negative loss of 80% of methyltransferase activity, leading to **focal hypomethylation in AML, especially at CGIs, shores and promoters** [790,791]. Other *DNMT3A* mutations cause both loss-of-function and gain-of-function effects [792]. In AML without *DNMT3A* mutations, upregulation of all methyltransferases leads to CGI hypermethylation, possibly as a consequence of rapid proliferation [793,794]. Certain oncoproteins, such as PML-RARA, can also recruit DNMT1 and DNMT3A to CGI promoters and promote hypermethylation [450].

Chromatin modifiers

Among chromatin modifiers, **several proteins related to the Polycomb group are mutated in AML**, and particularly in sAML progressing from MDS [694]. Mutations in *ASXL1* inhibit the recruitment of PRC2 by blocking its interaction with ASXL1 [795]. This leads to a global depletion of H3K27me3 that affects the HOXA cluster, essential for myeloid differentiation. However, even though loss of *ASXL1* alone disrupts hematopoiesis, it is insufficient to cause leukemia [796]. The PRC2 core subunits *EZH2* and *SUZ12* are frequently mutated in MDS and sAML, though rarely in primary AML, leading to reduced PRC2 histone methyltransferase activity [797,798]. This loss of function depletes H3K27me3 levels in leukemic cells, enabling the aberrant expression of genes involved in self-renewal, and promotes chemotherapy resistance in AML [799,800].

Mutations in *BCOR* and *BCORL1* are found in both AML and MDS, leading to either reduced expression or an inactive gene products [801–803]. The consequence of these mutations is the disruption of the PRC1.1 complex, which loses its repressive function, resulting in epigenetic rewiring and activation of HSPC signaling genes [804]. Since the expression of *BCOR* is higher than that of *BCORL1* in hematopoiesis, the effect of *BCOR1* mutations is more pronounced, which may explain why they are also more frequent. Mutations in *BCORL1* are often subclonal events following a prior *BCOR* mutation, indicating functional cooperation.

The histone methyltransferase *KMT2A*, also known as *MLL1*, is a partner in gene fusions created by chromosomal aberrations involving the q23 region of chromosome 11, among which t(9;11) is the most common [805,806]. The presence of these **11q23 translocations defines an AML subtype** with a unique transcriptional profile, often referred to as MLL-rearranged AML [767,807]. Many genes involved in transformation are highly expressed in these leukemias, such as *MEIS1* and HOX genes [807]; 43% of them overexpress *EVI1*, which promotes tumor growth and chemoresistance [808]. Wild type KMT2A mediates trimethylation of H3K4 via its SET domain, but this domain is lost in KMT2A fusions, in line with the fact that overexpression of *HOX* genes in MLL-rearranged AML does not involve H3K4me3 [809]. Rather, gene upregulation stems from the recruitment of the H3K79 methyltransferase DOT1L by KMT2A fusion partners like AF9 or AF10, resulting in ectopic H3K79 methylation [810,811]. Indeed, inactivation of DOT1L leads to downregulation of KMT2A fusion targets and suppression of the MLL-rearranged transcriptional signature [812]. Moreover, several KMT2A fusion partners, such as AF9 or ENL, are members of the SEC and contribute to misexpression via elongation dysregulation [813,814].

Genome organization

Mutations in members of the cohesin complex are present in more than 10% of the AML patients, but little is known about their effect on leukemogenesis and their prognostic impact is controversial [815–817]. Although loss-of-function cohesin mutations in other cancers cause chromosomal instability [818], they are not associated with cytogenetic abnormalities in AML [816]. Instead, they drive tumor progression by disrupting the formation of chromatin loops that ensure appropriate gene regulation, as shown by the unique transcriptional profile of AML with *STAG2* mutations [816]. Indeed, deletion of *STAG2* in HSCs does not cause any chromosomal aberrations, but it increases self-renewal and decreases differentiation with concomitant changes in genes associated with lineage specification [631]. These transcriptional changes seem to be caused by **loss of short-range interactions uniquely mediated by STAG2** in a CTCF-independent manner at regions bound by hematopoietic TFs [631], albeit other studies report alterations in TAD structure and compartments [819]. Similarly, even though complete loss of *SMC3* causes defects in sister chromatid separation, *SMC3* haploinsufficiency increases HSC self-renewal and cooperates with FLT3-ITD to induce leukemia in mice [820].

While most cohesin subunits exhibit mostly heterozygous missense mutations, STAG2 is frequently inactivated by truncated mutations of its single functional copy [182]. This discrepancy is possibly due to compensation by its paralog STAG1, as suggested by the synthetic lethality between STAG1 and STAG2 [821] and the overexpression of *STAG1* in *STAG2*-mutated AML [631]. By contrast, each core cohesin subunit is essential and cell viability would be compromised in their absence. Confirmatory evidence was provided by deletion experiments in HSCs showing that STAG1 compensates for STAG2 in higher order DNA organization and chromosomal segregation [631]. However, only *Stag2* KO increased HSC self-renewal, in line with prior reports of distinct roles of each protein in chromosome organization [204]. Perhaps for this reason, mutations in *STAG1* are exceedingly rare in AML.

4.4.2 Mutations in non-coding regions

Although cancer research has traditionally focused on coding regions, the importance of mutations in non-coding regions is becoming increasingly apparent. The development of technologies for the exploration of the epigenetic landscape, like ATAC-seq or ChIP-seq, has paved the way for a new wave of research that attempts to understand changes in regulatory regions associated with cancer. Special interest has been placed on super-enhancers, which are often acquired in the vicinity of oncogenes [537]. Common mechanisms for dysregulation of gene expression involving CREs are focal amplifications, translocations, TFBS-creating mutations and disruptions in TAD boundaries [822].

Expression of *EVI1*, encoded by *MECOM*, is normally restricted to HSCs in hematopoiesis, where it is critical for proliferation and repopulation capacity [823,824]. However, **a subset of ~8% of AMLs express high levels of *EVI1*,** which is an independent predictor for poor survival [825,826]. Overexpression of *EVI1* is thought to contribute to leukemogenesis by inhibiting myeloid differentiation [827] and promoting proliferation and survival of LSCs [824]. Among *EVI1*-expressing leukemias, roughly 20% carry chromosome 3q26 abnormalities, where *MECOM* is located, and another 20% exhibits MLL rearrangements [825]. Almost two thirds of 3q26-rearranged AMLs exhibit inv(3) or t(3;3) [828,829]. In the early 1990s, it was proposed that the activation of *EVI1* in AML with inv(3)/t(3;3) stemmed from the repositioning of an enhancer close to *RPN1* [830,831]. However, it was not until 2014 that the integration of epigenomics data demonstrated that ***EVI1* overexpression was driven by the hijacking of a translocated *GATA2* super-enhancer** [230,832]. It stands to reason that similar mechanisms operate in other 3q26-rearranged AMLs, but they have not been investigated so far. The activation of *EVI1* in the 60% of leukemias without 3q26 or 11q23 rearrangements remains an enigma.

Focal amplifications of a super-enhancer located 1.7 Mb downstream of *MYC* have been reported in 3% of AMLs, presumably increasing the concentration of TFs [833]. *MYC* is a proto-oncogene that is activated in the majority of cancers, leading to increased proliferation and tumor evasion [834]. The interaction between the *MYC* and its super-enhancer is dependent on the SWI/SNF chromatin remodeler, which facilitates the binding of TFs by displacing

nucleosomes [833]. The **formation of novel (super-)enhancers due to mutations in non-coding regions** has not been described in AML, but is well-known in T-ALL. Marc Mansour and colleagues reported somatic indels at a hotspot upstream of *TAL1*, resulting in the creation of a binding site for MYB, which can recruit CBP [538]. Acetylation of histone marks by CBP makes chromatin accessible to other hematopoietic TFs, leading to the formation of a super-enhancer that upregulates *TAL1* expression. A similar mechanism was later described for *LMO2*, an oncogenic driver of T-ALL that is normally silent in T-cells, but acquires an enhancer-forming insertion in 2% of T-ALL cases [835]. Furthermore, *LMO2* intronic mutations found in 5% of T-ALLs generate a neomorphic promoter, also leading to overexpression of this oncogene [836].

Disruptions in TAD boundaries can be caused by either genetic or epigenetic mechanisms. In AML with inv(3)/t(3;3), the rearrangement perturbs the usual boundaries of the TAD containing *MECOM*, leading to the formation of a chimeric TAD that also contains the *GATA2* super-enhancer [230]. An example of epigenetic disruption has been shown in glioma with *IDH1/2* mutations, in which the hypermethylation induced by the loss of IDH prevents the binding of CTCF to a TAD boundary that usually insulates *PDGFRA* from a nearby *FIP1L1* enhancer [228]. The existence of an identical mechanism in AML is an intriguing possibility, given the existence of multiple mutations that also affect DNA methylation, as mentioned above.

4.4.3 The role of epimutations in leukemogenesis

Much of the work discussed so far has attempted to relate epigenetic events to specific mutations, either in coding or non-coding regions. Nevertheless, heritable epigenetic changes can also occur in the absence of a underlying genetic lesion, a phenomenon designated as **epimutation** by Robin Holliday following his earlier work on methylation and gene silencing [837]. Epimutations can be categorized as primary or secondary depending on whether they are truly independent from DNA changes or are the indirect consequence of a mutation [838]. Thus, hypermethylation of a promoter in *TET2*-mutated AML would be a secondary event rather than a pure epigenetic event, which should be, by definition, reversible. Another example of secondary epimutation is the increased CTCF occupancy observed in AML, possibly due to global hypomethylation [839]. Furthermore, mutated NPM1 also causes the cytoplasmic mislocalization of CTCF [840].

Verified primary epimutations in cancer are rare because it is hard to exclude the possibility they result from a DNA change in *cis* or *trans* somewhere in the genome [841]. One of the few examples is the germline inactivation of *MLH1* by methylation, which predisposes to colorectal cancer [842]. Although not as well understood as the transmission of methylation marks, epigenetic inheritance of histone modifications has also been demonstrated over multiple cell divisions [843]. They are, therefore, also possible candidates for epimutations. An intriguing example in leukemia is the **hypermethylation of *CEBPA* in cases with mixed**

myeloid/lymphoid phenotype that exhibit a transcriptional program similar to AML with *CEBPA* DM [844]. Originally identified as *CEBPA*-silenced leukemias, it was later shown that aberrant methylation was not restricted to the *CEBPA* locus, but was in fact a genome-wide phenomenon [443]. In line with this, they were independently characterized as CIMP AMLs by a separate group [444]. Nevertheless, given their similarities with *CEBPA* DM AML, it can be argued that silencing of *CEBPA* is not only one in many changes, but a pivotal factor in the leukemogenic process. On the other hand, aberrant **hypermethylation of *DNMT3A* has been described in 40% of AML patients**, possibly acting as an epimutation with similar effects to loss-of-function genetic changes like R880H [845].

The mechanisms whereby these epimutations appear remain poorly understood. While they may be secondary to genetic lesions, the absence of recurrent mutations associated with these events points to alternative mechanisms. They could be an indirect consequence of mutations in non-coding regions regulating the expression of epigenetic "writers" or "erasers" (e.g. an enhancer of *TET2*), in which case it may be possible to detect them via gene expression profiling together with WGS. However, it is tempting to speculate that they are driven by random epigenetic variation followed by Darwinian selection, perhaps in the context of the widespread alterations that accompany lineage specification. In keeping with this hypothesis, substantial epigenetic diversity has been observed in AML patients, possibly driving cancer evolution [846].

5. SCOPE AND AIMS OF THIS THESIS

Despite tremendous progress in the last two decades, the understanding of epigenetic regulation in both health and malignant hematopoiesis remains incomplete. The studies included in this thesis aim to shed light on this burgeoning field by employing a combination of molecular biology and bioinformatics approaches that explore different aspects of epigenetic regulation. The thesis is divided into three sections:

1. Transcription factors in healthy hematopoiesis (chapter 2): in this section we investigated the roles of *CEBPA* in myeloid commitment and HSC maintenance. To this end, we studied a mouse model in which deletion of the hematopoietic +37 kb *Cebpa* enhancer leads to depletion of the LT-HSC compartment as well as a block in myeloid differentiation. Using transplantation experiments and transcriptomics techniques, both in bulk and in single cells, we determined whether the LT-HSC loss is cell-extrinsic or cell-intrisic.

2. Enhancer hijacking in AML (chapters 3-5): here we attempted to elucidate whether there is a common enhancer hijacking mechanism directing *EVI1* overexpression in all 3q26-rearranged AMLs and what components of transcriptional regulation are critical in this process. In **chapter 3**, we profiled a cohort of AML with atypical 3q26 rearrangements using genetic and epigenetic approaches to investigate their commonalities with the classical inv(3)/t(3;3) AML. Most of the translocations involved super-enhancers of genes active in myeloid development. This finding was further explored in **chapter 4**, where we generated a cell line model containing the t(3;8) rearrangement and dissected the translocated *MYC* super-enhancer to find critical regions involved in *EVI1* overexpression. We also examined how the same mechanism may operate in AMLs with other 3q26 rearrangements. In **chapter 5**, we conducted a CRISPR screen in the inv(3) cell line MUTZ3 to determine what sequences are uniquely essential in the rearranged *GATA2* enhancer, but not in the wild type allele. Some of the TFs binding to these sequences are amenable to pharmaceutical inhibition and thus constitute attractive points of therapeutic intervention.

3. Epigenetic dysregulation driving altered gene expression in AML (chapters 6-7): here we sought to identify somatic epigenetic events in AML and understand how they drive leukemogenesis. In **chapter 6**, we conducted a screen of genes with allele specific expression (ASE) in a cohort of 200 AML patients to detect changes in cis-regulatory elements. Since most genes are expressed in similar proportions from each allele in normal conditions, recurrent ASE of certain genes may unveil alterations in neighboring regulatory regions. In **chapter 7**, we profiled CEBPA-silenced/CIMP leukemias at the genetic and epigenetic level to understand the effects of methylation on cell identity and leukemogenesis. This study required the integration of DNA methylation, H3K27ac, H3K27me3, chromatin accessibility, CTCF binding and chromatin structure.

In **chapter 8**, I summarized the findings of this thesis and discussed their implications for our understanding of epigenetics in hematopoiesis.

REFERENCES

1. Khwaja, A. *et al.* Acute myeloid leukaemia. *Nat. Rev. Dis. Prim.* **2**, 16010 (2016).
2. The Cancer Genome Atlas Research Network *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–74 (2013).
3. Figueroa, M. E. *et al.* DNA Methylation Signatures Identify Biologically Distinct Subtypes in Acute Myeloid Leukemia. *Cancer Cell* **17**, 13–27 (2010).
4. Bhagwat, A. S., Lu, B. & Vakoc, C. R. Enhancer dysfunction in leukemia. *Blood* vol. 131 1795–1804 (2018).
5. Liggett, L. A. & Sankaran, V. G. Unraveling Hematopoiesis through the Lens of Genomics. *Cell* vol. 182 1384–1400 (2020).
6. Eberl, G., Colonna, M., Di Santo, J. P. & McKenzie, A. N. J. Innate lymphoid cells. Innate lymphoid cells: a new paradigm in immunology. *Science (80-.).* **348**, aaa6566 (2015).
7. Fliedner, T. M., Graessle, D., Paulsen, C. & Reimers, K. Structure and function of bone marrow hemopoiesis: mechanisms of response to ionizing radiation exposure. *Cancer Biother. Radiopharm.* **17**, 405–26 (2002).
8. Orkin, S. H. & Zon, L. I. Hematopoiesis: An Evolving Paradigm for Stem Cell Biology. *Cell* vol. 132 631–644 (2008).
9. Morrison, S. J., Uchida, N. & Weissman, I. L. The biology of hematopoietic stem cells. *Annual Review of Cell and Developmental Biology* vol. 11 35–71 (1995).
10. McCulloch, E. A. & Till, J. E. Perspectives on the properties of stem cells. *Nature Medicine* vol. 11 1026–1028 (2005).
11. Till, J. E. & McCulloch, E. A. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiat. Res.* **14**, 213–22 (1961).
12. Becker, A. J., McCulloch, E. A. & Till, J. E. Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. *Nature* **197**, 452–454 (1963).
13. Siminovitch, L., McCulloch, E. A. & Till, J. E. The distribution of colony-forming cells among spleen colonies. *J. Cell. Comp. Physiol.* **62**, 327–336 (1963).
14. Osawa, M., Hanada, K. I., Hamada, H. & Nakauchi, H. Long-term lymphohematopoietic reconstitution by a single CD34- low/negative hematopoietic stem cell. *Science (80-.).* **273**, 242–245 (1996).
15. Spangrude, G. J., Heimfeld, S. & Weissman, I. L. Purification and characterization of mouse hematopoietic stem cells. *Science (80-.).* **241**, 58–62 (1988).
16. Morrison, S. J. & Weissman, I. L. The long-term repopulating subset of hematopoietic stem cells is deterministic and isolatable by phenotype. *Immunity* **1**, 661–673 (1994).
17. Kondo, M. *et al.* Biology of hematopoietic stem cells and progenitors: Implications for clinical application. *Annual Review of Immunology* vol. 21 759–806 (2003).
18. Krause, D., Fackler, M., Civin, C. & May, W. CD34: structure, biology, and clinical utility. *Blood* **87**, 1–13 (1996).
19. AbuSamra, D. B. *et al.* Not just a marker: CD34 on human hematopoietic stem/progenitor cells dominates vascular selectin binding along with CD44. *Blood Adv.* **1**, 2799–2816 (2017).
20. Anjos-Afonso, F. *et al.* CD34- cells at the apex of the human hematopoietic stem cell hierarchy have distinctive cellular and molecular signatures. *Cell Stem Cell* **13**, 161–174 (2013).
21. Nakamura, Y. *et al.* Ex vivo generation of CD34+ cells from CD34- hematopoietic cells. *Blood* **94**, 4053–4059 (1999).
22. Anjos-Afonso, F. & Bonnet, D. Forgotten gems: Human CD34- hematopoietic stem cells. *Cell Cycle* **13**, 503–504 (2014).
23. Laurenti, E. & Göttgens, B. From haematopoietic stem cells to complex differentiation landscapes. *Nature* vol. 553 418–426 (2018).

24. Larochelle, A. *et al.* Human and rhesus macaque hematopoietic stem cells cannot be purified based only on SLAM family markers. *Blood* **117**, 1550–1554 (2011).

25. Randall, T. D., Lund, F. E., Howard, M. C. & Weissman, I. L. Expression of murine CD38 defines a population of long-term reconstituting hematopoietic stem cells. *Blood* **87**, 4057–4067 (1996).

26. Baum, C. M., Weissman, I. L., Tsukamoto, A. S., Buckle, A. M. & Peault, B. Isolation of a candidate human hematopoietic stem-cell population. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 2804–2808 (1992).

27. Manz, M. G., Miyamoto, T., Akashi, K. & Weissman, I. L. Prospective isolation of human clonogenic common myeloid progenitors. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 11872–11877 (2002).

28. Till, J. E., McCulloch, E. A. & Siminovitch, L. A stochastic model of stem cell proliferation, based on the growth of spleen colony-forming cells. *Proc. Natl. Acad. Sci. U. S. A.* **51**, 29–36 (1964).

29. Abkowitz, J. L., Catlin, S. N., McCallie, M. T. & Guttorp, P. Evidence that the number of hematopoietic stem cells per animal is conserved in mammals. *Blood* **100**, 2665–2667 (2002).

30. Seita, J. & Weissman, I. L. Hematopoietic stem cell: Self-renewal versus differentiation. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2**, 640–653 (2010).

31. Cheshier, S. H., Morrison, S. J., Liao, X. & Weissman, I. L. In vivo proliferation and cell cycle kinetics of long-term self-renewing hematopoietic stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 3120–3125 (1999).

32. Nakamura-Ishizu, A., Takizawa, H. & Suda, T. The analysis, roles and regulation of quiescence in hematopoietic stem cells. *Development (Cambridge)* vol. 141 4656–4666 (2014).

33. Brenet, F., Kermani, P., Spektor, R., Rafii, S. & Scandura, J. M. Tgfβ restores hematopoietic homeostasis after myelosuppressive chemotherapy. *J. Exp. Med.* **210**, 623–639 (2013).

34. Wilson, A. *et al.* Hematopoietic Stem Cells Reversibly Switch from Dormancy to Self-Renewal during Homeostasis and Repair. *Cell* **135**, 1118–1129 (2008).

35. Walter, D. *et al.* Exit from dormancy provokes DNA-damage-induced attrition in haematopoietic stem cells. *Nature* **520**, 549–552 (2015).

36. Anasetti, C. *et al.* Peripheral-blood stem cells versus bone marrow from unrelated donors. *N. Engl. J. Med.* **367**, 1487–96 (2012).

37. Wright, D. E. *et al.* Cyclophosphamide/granulocyte colony-stimulating factor causes selective mobilization of bone marrow hematopoietic stem cells into the blood after M phase of the cell cycle. *Blood* **97**, 2278–2285 (2001).

38. Bernitz, J. M., Daniel, M. G., Fstkchyan, Y. S. & Moore, K. Granulocyte colony-stimulating factor mobilizes dormant hematopoietic stem cells without proliferation in mice. *Blood* **129**, 1901–1912 (2017).

39. Lemischka, I. R., Raulet, D. H. & Mulligan, R. C. Developmental potential and dynamic behavior of hematopoietic stem cells. *Cell* **45**, 917–927 (1986).

40. Jordan, C. T. & Lemischka, I. R. Clonal and systemic analysis of long-term hematopoiesis in the mouse. *Genes Dev.* **4**, 220–232 (1990).

41. Abkowitz, J. L., Catlin, S. N. & Guttorp, P. Evidence that hematopoiesis may be a stochastic process in vivo. *Nat. Med.* **2**, 190–197 (1996).

42. Lemischka, I. R. Microenvironmental regulation of hematopoietic stem cells. *Stem Cells* **15**, 63–68 (1997).

43. Müller-Sieburg, C. E., Cho, R. H., Thoman, M., Adkins, B. & Sieburg, H. B. Deterministic regulation of hematopoietic stem cell self-renewal and differentiation. *Blood* **100**, 1302–1309 (2002).

44. Sieburg, H. B. *et al.* The hematopoietic stem compartment consists of a limited number of discrete stem cell subsets. *Blood* **107**, 2311–2316 (2006).

45. Dykstra, B. *et al.* Long-Term Propagation of Distinct Hematopoietic Differentiation Programs In Vivo. *Cell Stem Cell* **1**, 218–229 (2007).

46. Challen, G. A., Boles, N. C., Chambers, S. M. & Goodell, M. A. Distinct Hematopoietic Stem Cell Subtypes Are Differentially Regulated by TGF-β1. *Cell Stem Cell* **6**, 265–278 (2010).
47. Bendall, S. C. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science (80-.).* **332**, 687–696 (2011).
48. Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science (80-.).* **343**, 776–779 (2014).
49. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).
50. Paul, F. *et al.* Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **163**, 1663–1677 (2015).
51. Nestorowa, S. *et al.* A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128**, e20–e31 (2016).
52. Velten, L. *et al.* Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.* **19**, 271–281 (2017).
53. Notta, F. *et al.* Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science (80-.).* **351**, (2016).
54. Pellin, D. *et al.* A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nat. Commun.* **10**, (2019).
55. Pietras, E. M. *et al.* Functionally Distinct Subsets of Lineage-Biased Multipotent Progenitors Control Blood Production in Normal and Regenerative Conditions. *Cell Stem Cell* **17**, 35–46 (2015).
56. Giladi, A. *et al.* Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nat. Cell Biol.* **20**, 836–846 (2018).
57. Olsson, A. *et al.* Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature* **537**, 698–702 (2016).
58. Marinov, G. K. *et al.* From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Res.* **24**, 496–510 (2014).
59. Cheng, H., Zheng, Z. & Cheng, T. New paradigms on hematopoietic stem cell differentiation. *Protein and Cell* vol. 11 34–44 (2020).
60. Sun, J. *et al.* Clonal dynamics of native haematopoiesis. *Nature* **514**, 322–327 (2014).
61. Busch, K. *et al.* Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature* **518**, 542–546 (2015).
62. Sawai, C. M. *et al.* Hematopoietic Stem Cells Are the Major Source of Multilineage Hematopoiesis in Adult Animals. *Immunity* **45**, 597–609 (2016).
63. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science (80-.).* **367**, (2020).
64. Doulatov, S. *et al.* Revised map of the human progenitor hierarchy shows the origin of macrophages and dendritic cells in early lymphoid development. *Nat. Immunol.* **11**, 585–593 (2010).
65. Dexter, T. M., Allen, T. D. & Lajtha, L. G. Conditions controlling the proliferation of haemopoietic stem cells in vitro. *J. Cell. Physiol.* **91**, 335–344 (1977).
66. Schofield, R. The relationship between the spleen colony-forming cell and the haemopoietic stem cell. *Blood Cells* **4**, 7–25 (1978).
67. Morrison, S. J. & Scadden, D. T. The bone marrow niche for haematopoietic stem cells. *Nature* vol. 505 327–334 (2014).
68. Visnjic, D. *et al.* Hematopoiesis is severely altered in mice with an induced osteoblast deficiency. *Blood* **103**, 3258–3264 (2004).

69. Bruns, I. *et al.* Megakaryocytes regulate hematopoietic stem cell quiescence through CXCL4 secretion. *Nat. Med.* **20**, 1315–1320 (2014).

70. Pinho, S. & Frenette, P. S. Haematopoietic stem cell activity and interactions with the niche. *Nature Reviews Molecular Cell Biology* vol. 20 303–320 (2019).

71. Crane, G. M., Jeffery, E. & Morrison, S. J. Adult haematopoietic stem cell niches. *Nat. Rev. Immunol.* **17**, 573–590 (2017).

72. Moll, N. M. & Ransohoff, R. M. CXCL12 and CXCR4 in bone marrow physiology. *Expert Review of Hematology* vol. 3 315–322 (2010).

73. Ding, L., Saunders, T. L., Enikolopov, G. & Morrison, S. J. Endothelial and perivascular cells maintain haematopoietic stem cells. *Nature* **481**, 457–462 (2012).

74. Rossi, D. J. *et al.* Cell intrinsic alterations underlie hematopoietic stem cell aging. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 9194–9199 (2005).

75. Liang, Y., Van Zant, G. & Szilvassy, S. J. Effects of aging on the homing and engraftment of murine hematopoietic stem and progenitor cells. *Blood* **106**, 1479–1487 (2005).

76. Poulos, M. G. *et al.* Endothelial transplantation rejuvenates aged hematopoietic stem cell function. *J. Clin. Invest.* **127**, 4163–4178 (2017).

77. Raaijmakers, M. H. G. P. *et al.* Bone progenitor dysfunction induces myelodysplasia and secondary leukaemia. *Nature* **464**, 852–857 (2010).

78. Schepers, K. *et al.* Myeloproliferative neoplasia remodels the endosteal bone marrow niche into a self-reinforcing leukemic niche. *Cell Stem Cell* **13**, 285–299 (2013).

79. Hanoun, M. *et al.* Acute myelogenous leukemia-induced sympathetic neuropathy promotes malignancy in an altered hematopoietic stem cell Niche. *Cell Stem Cell* **15**, 365–375 (2014).

80. Waddington, C. H. The epigenotype. *Endeavour* 18–20 (1942) doi:10.1093/ije/dyr184.

81. Waddington, C. H. & Kacser H. *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology*. *The Strategy of the Genes* (Allen & Unwin, 1957). doi:10.4324/9781315765471.

82. Ferrell, J. E. Bistability, bifurcations, and Waddington's epigenetic landscape. *Current Biology* vol. 22 R458–R466 (2012).

83. Moris, N., Pina, C. & Arias, A. M. Transition states and cell fate decisions in epigenetic landscapes. *Nature Reviews Genetics* vol. 17 693–703 (2016).

84. McClintock, B. Controlling elements and the gene. *Cold Spring Harb. Symp. Quant. Biol.* **21**, 197–216 (1956).

85. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology* vol. 3 318–356 (1961).

86. Deichmann, U. Epigenetics: The origins and evolution of a fashionable topic. *Dev. Biol.* **416**, 249–254 (2016).

87. Rodrigues, C. P., Shvedunova, M. & Akhtar, A. Epigenetic Regulators as the Gatekeepers of Hematopoiesis. *Trends in Genetics* vol. 37 125–142 (2021).

88. Abdellah, Z. *et al.* Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).

89. Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* **5**, 1000598 (2009).

90. Sainsbury, S., Bernecky, C. & Cramer, P. Structural basis of transcription initiation by RNA polymerase II. *Nature Reviews Molecular Cell Biology* vol. 16 129–143 (2015).

91. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: Emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics* vol. 13 233–245 (2012).

92. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology* vol. 19 621–637 (2018).

93. Spitz, F. & Furlong, E. E. M. Transcription factors: From enhancer binding to developmental control. *Nature Reviews Genetics* vol. 13 613–626 (2012).
94. Ong, C. & Corces, V. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.* **12**, 283–93 (2011).
95. Segert, J. A., Gisselbrecht, S. S. & Bulyk, M. L. Transcriptional Silencers: Driving Gene Expression with the Brakes On. *Trends in Genetics* vol. 37 514–527 (2021).
96. Ghirlando, R. & Felsenfeld, G. CTCF: Making the right connections. *Genes and Development* vol. 30 881–891 (2016).
97. Li, G. & Reinberg, D. Chromatin higher-order structures and gene regulation. *Current Opinion in Genetics and Development* vol. 21 175–186 (2011).
98. Allshire, R. C. & Madhani, H. D. Ten principles of heterochromatin formation and function. *Nature Reviews Molecular Cell Biology* vol. 19 229–244 (2018).
99. Li, W., Notani, D. & Rosenfeld, M. G. Enhancers as non-coding RNA transcription units: Recent insights and future perspectives. *Nature Reviews Genetics* vol. 17 207–223 (2016).
100. Olins, A. L. & Olins, D. E. Spheroid chromatin units (v bodies). *Science (80-.).* **183**, 330–332 (1974).
101. Kornberg, R. D. & Thomas, J. O. Chromatin structure: Oligomers of the histones. *Science (80-.).* **184**, 865–868 (1974).
102. Li, G. & Zhu, P. Structure and organization of chromatin fiber in the nucleus. *FEBS Letters* vol. 589 2893–2904 (2015).
103. Fyodorov, D. V., Zhou, B. R., Skoultchi, A. I. & Bai, Y. Emerging roles of linker histones in regulating chromatin structure and function. *Nat. Rev. Mol. Cell Biol.* **19**, 192–206 (2018).
104. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006).
105. Kornberg, R. D. & Lorch, Y. Primary Role of the Nucleosome. *Molecular Cell* vol. 79 371–375 (2020).
106. Finch, J. T. & Klug, A. Solenoidal model for superstructure in chromatin. *Proc. Natl. Acad. Sci. U. S. A.* **73**, 1897–1901 (1976).
107. Fussner, E. *et al.* Open and closed domains in the mouse genome are configured as 10-nm chromatin fibres. *EMBO Rep.* **13**, 992–996 (2012).
108. Baldi, S., Korber, P. & Becker, P. B. Beads on a string—nucleosome array arrangements and folding of the chromatin fiber. *Nature Structural and Molecular Biology* vol. 27 109–118 (2020).
109. Ricci, M. A., Manzo, C., García-Parajo, M. F., Lakadamyali, M. & Cosma, M. P. Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. *Cell* **160**, 1145–1158 (2015).
110. Gross, D. S. & Garrard, W. T. Nuclease hypersensitive sites in chromatin. *Annual Review of Biochemistry* vol. 57 159–197 (1988).
111. Lorch, Y., LaPointe, J. W. & Kornberg, R. D. Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell* **49**, 203–210 (1987).
112. Belotserkovskaya, R. *et al.* FACT facilitates transcription-dependent nucleosome alteration. *Science (80-.).* **301**, 1090–1093 (2003).
113. Kulaeva, O. I., Hsieh, F. K., Chang, H. W., Luse, D. S. & Studitsky, V. M. Mechanism of transcription through a nucleosome by RNA polymerase II. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* vol. 1829 76–83 (2013).
114. Lee, C. K., Shibata, Y., Rao, B., Strahl, B. D. & Lieb, J. D. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat. Genet.* **36**, 900–905 (2004).
115. Galas, D. J. & Schmitz, A. DNAase footprinting a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.* **5**, 3157–3170 (1978).

116. Wu, C., Wong, Y. C. & Elgin, S. C. R. The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. *Cell* **16**, 807–14 (1979).

117. Wu, C. The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. *Nature* **286**, 854–60 (1980).

118. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods* **6**, 283–9 (2009).

119. Sung, M.-H., Baek, S. & Hager, G. L. Genome-wide footprinting: ready for prime time? *Nat. Methods* **13**, 222–228 (2016).

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

121. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).

122. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science (80-.).* **362**, (2018).

123. Karabacak Calviello, A., Hirsekorn, A., Wurmus, R., Yusuf, D. & Ohler, U. Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling. *Genome Biol.* **20**, 1–13 (2019).

124. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

125. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: A hitchhiker's guide to ATAC-seq data analysis. *Genome Biology* vol. 21 1–16 (2020).

126. Henikoff, S. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nature Reviews Genetics* vol. 9 15–26 (2008).

127. Struhl, K. & Segal, E. Determinants of nucleosome positioning. *Nature Structural and Molecular Biology* vol. 20 267–273 (2013).

128. Thåström, A. *et al.* Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *J. Mol. Biol.* **288**, 213–229 (1999).

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

130. Lyer, V. & Struhl, K. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.* **14**, 2570–2579 (1995).

131. Schones, D. E. *et al.* Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell* **132**, 887–898 (2008).

132. Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474**, 516–522 (2011).

133. Zhang, Y. *et al.* Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nat. Struct. Mol. Biol.* **16**, 847–852 (2009).

134. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362–366 (2009).

135. Clapier, C. R., Iwasa, J., Cairns, B. R. & Peterson, C. L. Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nature Reviews Molecular Cell Biology* vol. 18 407–422 (2017).

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

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

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

139. Boyes, J., Omichinski, J., Clark, D., Pikaart, M. & Felsenfeld, G. Perturbation of nucleosome structure by the erythroid transcription factor GATA-1. *J. Mol. Biol.* **279**, 529–544 (1998).

140. Hu, G. *et al.* Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. *Genome Res.* **21**, 1650–1658 (2011).

141. Suganuma, T. & Workman, J. L. Signals and combinatorial functions of histone modifications. *Annu. Rev. Biochem.* **80**, 473–499 (2011).

142. Hassan, A. H. *et al.* Function and selectivity of bromodomains in anchoring chromatin-modifying complexes to promoter nucleosomes. *Cell* **111**, 369–379 (2002).

143. Hassan, A. H., Neely, K. E. & Workman, J. L. Histone acetyltransferase complexes stabilize SWI/SNF binding to promoter nucleosomes. *Cell* **104**, 817–827 (2001).

144. Hassan, A. H., Awad, S. & Prochasson, P. The Swi2/Snf2 bromodomain is required for the displacement of SAGA and the octamer transfer of SAGA-+acetylated nucleosomes. *J. Biol. Chem.* **281**, 18126–18134 (2006).

145. Cosma, M. P., Tanaka, T. & Nasmyth, K. Ordered Recruitment of Transcription and Chromatin Remodeling Factors to a Cell Cycle– and Developmentally Regulated Promoter. *Cell* **97**, 299–311 (1999).

146. Francis, N. J., Saurin, A. J., Shao, Z. & Kingston, R. E. Reconstitution of a functional core polycomb repressive complex. *Mol. Cell* **8**, 545–556 (2001).

147. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science (80-.).* **295**, 1306–1311 (2002).

148. Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat. Rev. Genet.* **19**, 789–800 (2018).

149. Wang, H., Han, M. & Qi, L. S. Engineering 3D genome organization. *Nature Reviews Genetics* vol. 22 343–360 (2021).

150. Schardin, M., Cremer, T., Hager, H. D. & Lang, M. Specific staining of human chromosomes in Chinese hamster x man hybrid cell lines demonstrates interphase chromosome territories. *Hum. Genet.* **71**, 281–287 (1985).

151. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (80-.).* **326**, 289–293 (2009).

152. Quinodoz, S. A. *et al.* Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* (2018) doi:10.1016/J.CELL.2018.05.024.

153. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. (2014) doi:10.1016/j.cell.2014.11.021.

154. Cullen, K. E., Kladde, M. P. & Seyfred, M. A. Interaction between transcription regulatory regions of prolactin chromatin. *Science (80-.).* **261**, 203–206 (1993).

155. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* **38**, 1348–54 (2006).

156. Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).

157. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458–72 (2012).

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

159. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* (2012) doi:10.1038/nature11049.

160. Dowen, J. M. *et al.* Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374–387 (2014).

161. Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163**, 1611–1627 (2015).

162. Benner, C., Isoda, T. & Murre, C. New roles for DNA cytosine modification, eRNA, anchors, and superanchors in developing B cell progenitors. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 12776–12781 (2015).

163. Pal, K., Forcato, M. & Ferrari, F. Hi-C analysis: from data generation to integration. *Biophysical Reviews* vol. 11 67–78 (2019).

164. Phillips-Cremins, J. E. *et al.* Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281–1295 (2013).

165. Tan, L., Xing, D., Chang, C. H., Li, H. & Xie, X. S. Three-dimensional genome structures of single diploid human cells. *Science (80-.).* **361**, 924–928 (2018).

166. Stevens, T. J. *et al.* 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59–64 (2017).

167. Schoenfelder, S. & Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics* vol. 20 437–455 (2019).

168. Ptashne, M. Gene regulation by proteins acting nearby and at a distance. *Nature* **322**, 697–701 (1986).

169. Irani, M. H., Orosz, L. & Adhya, S. A control element within a structural gene: the gal operon of Escherichia coli. *Cell* **32**, 783–788 (1983).

170. Dunn, T. M., Hahn, S., Ogden, S. & Schleif, R. F. An operator at -280 base pairs that is required for repression of araBAD operon promoter: addition of DNA helical turns between the operator and promoter cyclically hinders repression. *Proc. Natl. Acad. Sci. U. S. A.* **81**, 5017–5020 (1984).

171. Müller, H. P., Sogo, J. M. & Schaffner, W. An enhancer stimulates transcription in Trans when attached to the promoter via a protein bridge. *Cell* **58**, 767–777 (1989).

172. Carter, D., Chakalova, L., Osborne, C. S., Dai, Y. feng & Fraser, P. Long-range chromatin regulatory interactions in vivo. *Nat. Genet.* **32**, 623–626 (2002).

173. Deng, W. *et al.* Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**, 1233–1244 (2012).

174. Szabo, Q. *et al.* Regulation of single-cell genome organization into TADs and chromatin nanodomains. *Nat. Genet.* **52**, 1151–1157 (2020).

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

176. Pant, V. *et al.* Mutation of a Single CTCF Target Site within the H19 Imprinting Control Region Leads to Loss of Igf2 Imprinting and Complex Patterns of De Novo Methylation upon Maternal Inheritance. *Mol. Cell. Biol.* **24**, 3497 (2004).

177. Aljahani, A. *et al.* Analysis of sub-kilobase chromatin topology reveals nano-scale regulatory interactions with variable dependence on cohesin and CTCF. *Nat. Commun.* **13**, 1–13 (2022).

178. de Wit, E. *et al.* CTCF Binding Polarity Determines Chromatin Looping. *Mol. Cell* **60**, 676–684 (2015).

179. Franke, M. *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).

180. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).

181. Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science (80-.).* **351**, 1454–1458 (2016).

182. Waldman, T. Emerging themes in cohesin cancer biology. *Nature Reviews Cancer* vol. 20 504–515 (2020).

183. Jeppsson, K., Kanno, T., Shirahige, K. & Sjögren, C. The maintenance of chromosome structure: Positioning and functioning of SMC complexes. *Nature Reviews Molecular Cell Biology* vol. 15 601–614 (2014).

184. Birkenbihl, R. P. & Subramani, S. Cloning and characterization of rad21 an essential gene of Schizosaccharomyces pombe involved in DNA double-strand-break repair. *Nucleic Acids Res.* **20**, 6605–6611 (1992).

185. Guacci, V., Koshland, D. & Strunnikov, A. A direct link between sister chromatid cohesion and chromosome condensation revealed through the analysis of MCD1 in S. cerevisiae. *Cell* **91**, 47–57 (1997).

186. Michaelis, C., Ciosk, R. & Nasmyth, K. Cohesins: Chromosomal proteins that prevent premature separation of sister chromatids. *Cell* **91**, 35–45 (1997).

187. Hirano, T., Kobayashi, R. & Hirano, M. Condensins, chromosome condensation protein complexes containing XCAP-C, XCAP-E and a Xenopus homolog of the Drosophila Barren protein. *Cell* **89**, 511–521 (1997).

188. Rollins, R. A., Korom, M., Aulner, N., Martens, A. & Dorsett, D. Drosophila Nipped-B Protein Supports Sister Chromatid Cohesion and Opposes the Stromalin/Scc3 Cohesion Factor To Facilitate Long-Range Activation of the cut Gene. *Mol. Cell. Biol.* **24**, 3100–3111 (2004).

189. Horsfield, J. A. *et al.* Cohesin-dependent regulation of Runx genes. *Development* **134**, 2639–2649 (2007).

190. Wendt, K. S. *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451**, 796–801 (2008).

191. Parelho, V. *et al.* Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms. *Cell* **132**, 422–433 (2008).

192. Hadjur, S. *et al.* Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature* **460**, 410–413 (2009).

193. Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430–435 (2010).

194. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* **15**, 2038–2049 (2016).

195. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci.* **112**, E6456–E6465 (2015).

196. Alipour, E. & Marko, J. F. Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res.* **40**, 11202–11212 (2012).

197. Davidson, I. F. & Peters, J. M. Genome folding through loop extrusion by SMC complexes. *Nature Reviews Molecular Cell Biology* vol. 22 445–464 (2021).

198. Davidson, I. F. *et al.* Rapid movement and transcriptional re-localization of human cohesin on DNA. *EMBO J.* **35**, 2671–2685 (2016).

199. Ganji, M. *et al.* Real-time imaging of DNA loop extrusion by condensin. *Science (80-.).* eaar7831 (2018) doi:10.1126/science.aar7831.

200. Gibcus, J. H. *et al.* A pathway for mitotic chromosome formation. *Science (80-.).* **359**, (2018).

201. Nagano, T. *et al.* Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61–67 (2017).

202. Zhang, H. *et al.* Chromatin structure dynamics during the mitosis-to-G1 phase transition. *Nature* **576**, 158–162 (2019).

203. Vian, L. *et al.* The Energetics and Physiological Impact of Cohesin Extrusion. *Cell* **173**, 1165-1178.e20 (2018).

204. Kojic, A. *et al.* Distinct roles of cohesin-SA1 and cohesin-SA2 in 3D chromosome organization. *Nat. Struct. Mol. Biol.* **25**, 496–504 (2018).

205. Weintraub, A. S. *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* **171**, 1573-1588.e28 (2017).

206. Beagan, J. A. *et al.* YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res.* **27**, 1139–1152 (2017).

207. Schwalie, P. C. *et al.* Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes. *Genome Biol.* **14**, R148 (2013).

208. El Khattabi, L. *et al.* A Pliable Mediator Acts as a Functional Rather Than an Architectural Bridge between Promoters and Enhancers. *Cell* **178**, 1145-1158.e20 (2019).

209. Jaeger, M. G. *et al.* Selective Mediator dependence of cell-type-specifying transcription. *Nat. Genet.* **52**, 719–727 (2020).

210. Schmidt, D. *et al.* A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res.* **20**, 578 (2010).

211. Abboud, N. *et al.* A cohesin-OCT4 complex mediates Sox enhancers to prime an early embryonic lineage. *Nat. Commun.* **6**, 1–14 (2015).

212. Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944.e22 (2017).

213. Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320.e24 (2017).

214. Haarhuis, J. H. I. *et al.* The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell* **169**, 693-707.e14 (2017).

215. Kim, S. & Shendure, J. Mechanisms of Interplay between Transcription Factors and the 3D Genome. *Molecular Cell* vol. 76 306–319 (2019).

216. Love, P. E., Warzecha, C. & Li, L. Q. Ldb1 complexes: The new master regulators of erythroid gene transcription. *Trends Genet.* **30**, 1–9 (2014).

217. Abramo, K. *et al.* A chromosome folding intermediate at the condensin-to-cohesin transition during telophase. *Nat. Cell Biol.* **21**, 1393–1402 (2019).

218. Moore, J. M. *et al.* Loss of maternal CTCF is associated with peri-implantation lethality of CtCf null embryos. *PLoS One* **7**, e34915 (2012).

219. Gregor, A. *et al.* De novo mutations in the genome organizer CTCF cause intellectual disability. *Am. J. Hum. Genet.* **93**, 124–131 (2013).

220. Grinfeld, J. *et al.* Classification and Personalized Prognosis in Myeloproliferative Neoplasms. *N. Engl. J. Med.* **379**, 1416–1430 (2018).

221. Alexander, T. B. *et al.* The genetic basis and cell of origin of mixed phenotype acute leukaemia. *Nature* **562**, 373–406 (2018).

222. Gabriele, M. *et al.* YY1 Haploinsufficiency Causes an Intellectual Disability Syndrome Featuring Transcriptional and Chromatin Dysfunction. *Am. J. Hum. Genet.* **100**, 907 (2017).

223. Agarwal, N. & Theodorescu, D. The Role of Transcription Factor YY1 in the Biology of Cancer. doi:10.1615/CritRevOncog.2017021071.

224. Piché, J., Van Vliet, P. P., Pucéat, M. & Andelfinger, G. The expanding phenotypes of cohesinopathies: one ring to rule them all! *Cell Cycle* vol. 18 2828–2848 (2019).

225. Kon, A. *et al.* Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. *Nat. Genet. 2013 4510* **45**, 1232–1237 (2013).

226. Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* **47**, 818–821 (2015).

227. Guo, Y. A. *et al.* Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. *Nat. Commun.* **9**, (2018).

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

229. Pinoli, P., Stamoulakatou, E., Nguyen, A. P., Martínez, M. R. & Ceri, S. Pan-cancer analysis of somatic mutations and epigenetic alterations in insulated neighbourhood boundaries. *PLoS One* **15**, e0227180 (2020).

230. Gröschel, S. *et al.* A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in Leukemia. *Cell* **157**, 369–381 (2014).

231. Canela, A. *et al.* Genome Organization Drives Chromosome Fragility. *Cell* **170**, 507-521.e18 (2017).

232. Gómez-Herreros, F. DNA Double Strand Breaks and Chromosomal Translocations Induced by DNA Topoisomerase II. *Frontiers in Molecular Biosciences* vol. 6 141 (2019).

233. Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W. & Richmond, T. J. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol.* **319**, 1097–1113 (2002).

234. Luger, K., Dechassa, M. L. & Tremethick, D. J. New insights into nucleosome and chromatin structure: An ordered state or a disordered affair? *Nature Reviews Molecular Cell Biology* vol. 13 436–447 (2012).

235. Pepenella, S., Murphy, K. J. & Hayes, J. J. Intra- and inter-nucleosome interactions of the core histone tail domains in higher-order chromatin structure. *Chromosoma* vol. 123 3–13 (2014).

236. Ghoneim, M., Fuchs, H. A. & Musselman, C. A. Histone Tail Conformations: A Fuzzy Affair with DNA. *Trends Biochem. Sci.* **46**, 564–578 (2021).

237. Ausio, J., Dong, F. & van Holde, K. E. Use of selectively trypsinized nucleosome core particles to analyze the role of the histone 'tails' in the stabilization of the nucleosome. *J. Mol. Biol.* **206**, 451–463 (1989).

238. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* vol. 403 41–45 (2000).

239. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* vol. 293 1074–1080 (2001).

240. Allfrey, V. G., Faulkner, R. & Mirsky, A. E. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proc. Natl. Acad. Sci. United States* **51**, 786–794 (1964).

241. Freitas, M. A., Sklenar, A. R. & Parthun, M. R. Application of mass spectrometry to the identification and quantification of histone post-translational modifications. *J. Cell. Biochem.* **92**, 691–700 (2004).

242. Kouzarides, T. Chromatin Modifications and Their Function. *Cell* **128**, 693–705 (2007).

243. Tessarz, P. & Kouzarides, T. Histone core modifications regulating nucleosome structure and dynamics. *Nat. Rev. Mol. Cell Biol. 2014 1511* **15**, 703–708 (2014).

244. Demetriadou, C., Koufaris, C. & Kirmizis, A. Histone N-alpha terminal modifications: genome regulation at the tip of the tail. *Epigenetics Chromatin 2020 131* **13**, 1–13 (2020).

245. Dhalluin, C. *et al.* Structure and ligand of a histone acetyltransferase bromodomain. *Nature* **399**, 491–496 (1999).

246. Brownell, J. E. *et al.* Tetrahymena histone acetyltransferase A: A homolog to yeast Gcn5p linking histone acetylation to gene activation. *Cell* **84**, 843–851 (1996).

247. Kuo, M. H. *et al.* Transcription-linked acetylation by Gcn5p of histones H3 and H4 at specific lysines. *Nature* **383**, 269–72 (1996).

248. Taunton, J., Hassig, C. A. & Schreiber, S. L. A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p. *Science (80-.).* **272**, 408–411 (1996).

249. Inoue, A. & Fujimoto, D. Enzymatic deacetylation of histone. *Biochem. Biophys. Res. Commun.* **36**, 146–150 (1969).

250. Sassone-Corsi, P. *et al.* Requirement of Rsk-2 for epidermal growth factor-activated phosphorylation of histone H3. *Science (80-.).* **285**, 886–891 (1999).

251. Hsu, J. Y. *et al.* Mitotic phosphorylation of histone H3 is governed by Ipl1/aurora kinase and Glc7/PP1 phosphatase in budding yeast and nematodes. *Cell* **102**, 279–291 (2000).

252. Rea, S. *et al.* Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* **406**, 593–599 (2000).

253. Chen, D. *et al.* Regulation of transcription by a protein methyltransferase. *Science (80-.).* **284**, 2174–2177 (1999).

254. Shi, Y. *et al.* Histone Demethylation Mediated by the Nuclear Amine Oxidase Homolog LSD1. *Cell* **119**, 941–953 (2004).

255. Paik, W. K. & Kim, S. Enzymatic demethylation of calf thymus histones. *Biochem. Biophys. Res. Commun.* **51**, 781–788 (1973).

256. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nature Reviews Genetics* vol. 17 487–500 (2016).

257. Turner, B. M. Defining an epigenetic code. *Nat. Cell Biol.* **9**, 2–6 (2007).

258. Zhao, S., Allis, C. D. & Wang, G. G. The language of chromatin modification in human cancers. *Nature Reviews Cancer* vol. 21 413–430 (2021).

259. Zhao, Y. & Garcia, B. A. Comprehensive Catalog of Currently Documented Histone Modifications. *Cold Spring Harb. Perspect. Biol.* **7**, a025064 (2015).

260. Marmorstein, R. & Zhou, M. M. Writers and readers of histone acetylation: Structure, mechanism, and inhibition. *Cold Spring Harb. Perspect. Biol.* **6**, a018762 (2014).

261. Black, J. C., Van Rechem, C. & Whetstine, J. R. Histone Lysine Methylation Dynamics: Establishment, Regulation, and Biological Impact. *Molecular Cell* vol. 48 491–507 (2012).

262. Blanc, R. S. & Richard, S. Arginine Methylation: The Coming of Age. *Mol. Cell* **65**, 8–24 (2017).

263. Rossetto, D., Avvakumov, N. & Côté, J. Histone phosphorylation: A chromatin modification involved in diverse nuclear events. *Epigenetics* **7**, 1098 (2012).

264. Smeenk, G. & Mailand, N. Writers, readers, and erasers of histone ubiquitylation in DNA double-strand break repair. *Frontiers in Genetics* vol. 7 122 (2016).

265. Ryu, H.-Y. & Hochstrasser, M. Histone sumoylation and chromatin dynamics. *Nucleic Acids Res.* **49**, 6043–6052 (2021).

266. Hottiger, M. O. Nuclear ADP-ribosylation and its role in chromatin plasticity, cell differentiation, and epigenetics. *Annual Review of Biochemistry* vol. 84 227–263 (2015).

267. Sakabe, K., Wang, Z. & Hart, G. W. β-N-acetylglucosamine (O-GlcNAc) is part of the histone code. *Proc. Natl. Acad. Sci.* **107**, 19915–19920 (2010).

268. Toleman, C. A. *et al.* Structural basis of O-GlcNAc recognition by mammalian 14-3-3 proteins. *Proc. Natl. Acad. Sci.* **115**, 5956–5961 (2018).

269. Xu, Y. M., Du, J. Y. & Lau, A. T. Y. Posttranslational modifications of human histone H3: An update. *Proteomics* vol. 14 2047–2060 (2014).

270. Wan, J., Liu, H., Chu, J. & Zhang, H. Functions and mechanisms of lysine crotonylation. *J. Cell. Mol. Med.* **23**, 7163 (2019).

271. Bernstein, B. E. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125**, 315–326 (2006).

272. Atlasi, Y. & Stunnenberg, H. G. The interplay of epigenetic marks during stem cell differentiation and development. *Nature Reviews Genetics* vol. 18 643–658 (2017).

273. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nature Reviews Molecular Cell Biology* vol. 16 144–154 (2015).

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

275. Hyun, K., Jeon, J., Park, K. & Kim, J. Writing, erasing and reading histone lysine methylations. *Experimental and Molecular Medicine* vol. 49 e324–e324 (2017).

276. Guccione, E. & Richard, S. The regulation, functions and clinical relevance of arginine methylation. *Nature Reviews Molecular Cell Biology* vol. 20 642–657 (2019).

277. Greer, E. L. & Shi, Y. Histone methylation: A dynamic mark in health, disease and inheritance. *Nature Reviews Genetics* vol. 13 343–357 (2012).

278. Dawson, M. A. The cancer epigenome: Concepts, challenges, and therapeutic opportunities. *Science (80-.).* **355**, 1147–1152 (2017).

279. Furey, T. S. ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews Genetics* vol. 13 840–852 (2012).

280. Solomon, M. J., Larsen, P. L. & Varshavsky, A. Mapping protein DNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene. *Cell* vol. 53 937–947 (1988).

281. Solomon, M. J. & Varshavsky, A. Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. *Proc. Natl. Acad. Sci. U. S. A.* **82**, 6470–4 (1985).

282. Kuo, M. H. & Allis, C. D. In vivo cross-linking and immunoprecipitation for studying dynamic Protein:DNA associations in a chromatin environment. *Methods A Companion to Methods Enzymol.* **19**, 425–433 (1999).

283. Hecht, A., Strahl-Bolsinger, S. & Grunstein, M. Spreading of transcriptional repressor SIR3 from telomeric heterochromatin. *Nature* **383**, 92–96 (1996).

284. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science (80-.).* **290**, 2306–2309 (2000).

285. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–502 (2007).

286. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**, 823–837 (2007).

287. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).

288. Feingold, E. A. *et al.* The ENCODE (ENCyclopedia of DNA Elements) Project. *Science* vol. 306 636–640 (2004).

289. Martens, J. H. A. & Stunnenberg, H. G. BLUEPRINT: mapping human blood cell epigenomes. *Haematologica* **98**, 1487–9 (2013).

290. Rotem, A. *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol. 2015 3311* **33**, 1165–1172 (2015).

291. Baranello, L., Kouzine, F., Sanford, S. & Levens, D. ChIP bias as a function of cross-linking time. *Chromosom. Res.* **24**, 175–181 (2016).

292. Meyer, C. A. & Liu, X. S. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics* vol. 15 709–721 (2014).

293. Kasinathan, S., Orsi, G. A., Zentner, G. E., Ahmad, K. & Henikoff, S. High-resolution mapping of transcription factor binding sites on native chromatin. *Nat. Methods* **11**, 203–209 (2014).

294. Das, P. M., Ramachandran, K., VanWert, J. & Singal, R. Chromatin immunoprecipitation assay. *BioTechniques* vol. 37 961–969 (2004).

295. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6**, (2017).

296. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun. 2019 101* **10**, 1–10 (2019).

297. Phillips, D. M. The presence of acetyl groups of histones. *Biochem. J.* **87**, 258–63 (1963).

298. Struhl, K. Histone acetylation and transcriptional regulatory mechanisms. *Genes and Development* vol. 12 599–606 (1998).

299. Hong, L., Schroth, G. P., Matthews, H. R., Yau, P. & Bradbury, E. M. Studies of the DNA binding properties of histone H4 amino terminus. Thermal denaturation studies reveal that acetylation markedly reduces the binding constant of the H4 "tail" to DNA. *J. Biol. Chem.* **268**, 305–314 (1993).

300. Lee, D. Y., Hayes, J. J., Pruss, D. & Wolffe, A. P. A positive role for histone acetylation in transcription factor access to nucleosomal DNA. *Cell* **72**, 73–84 (1993).

301. Filippakopoulos, P. *et al.* Selective inhibition of BET bromodomains. *Nat. 2010 4687327* **468**, 1067–1073 (2010).

302. Simone, C. & Peserico, A. Physical and functional HAT/HDAC interplay regulates protein acetylation balance. *J. Biomed. Biotechnol.* **2011**, (2011).

303. Zentner, G. E. & Henikoff, S. Regulation of nucleosome dynamics by histone modifications. *Nature Structural and Molecular Biology* vol. 20 259–266 (2013).

304. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet. 2008 407* **40**, 897–903 (2008).

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

306. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci.* **107**, 21931–21936 (2010).

307. Karmodiya, K., Krebs, A. R., Oulad-Abdelghani, M., Kimura, H. & Tora, L. H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics 2012 131* **13**, 1–18 (2012).

308. Pradeepa, M. M. *et al.* Histone H3 globular domain acetylation identifies a new class of enhancers. *Nat. Genet.* **48**, 681–686 (2016).

309. Taylor, G. C. A., Eskeland, R., Hekimoglu-Balkan, B., Pradeepa, M. M. & Bickmore, W. A. H4K16 acetylation marks active genes and enhancers of embryonic stem cells, but does not alter chromatin compaction. *Genome Res.* **23**, 2053–2065 (2013).

310. Di Cerbo, V. *et al.* Acetylation of histone H3 at lysine 64 regulates nucleosome dynamics and facilitates transcription. *Elife* **2014**, (2014).

311. Tropberger, P. *et al.* Regulation of transcription through acetylation of H3K122 on the lateral surface of the histone octamer. *Cell* **152**, 859–872 (2013).

312. Shogren-Knaak, M. *et al.* Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science (80-.).* **311**, 844–847 (2006).

313. Murray, K. The Occurrence of ϵ-N-Methyl Lysine in Histones. *Biochemistry* **3**, 10–15 (1964).

314. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet. 2007 393* **39**, 311–318 (2007).

315. Peters, A. H. F. M. *et al.* Partitioning and Plasticity of Repressive Histone Methylation States in Mammalian Chromatin. *Mol. Cell* **12**, 1577–1589 (2003).

316. Chantalat, S. *et al.* Histone H3 trimethylation at lysine 36 is associated with constitutive and facultative heterochromatin. *Genome Res.* **21**, 1426 (2011).

317. Mills, A. A. Throwing the cancer switch: Reciprocal roles of polycomb and trithorax proteins. *Nature Reviews Cancer* vol. 10 669–682 (2010).

318. Feng, Q. *et al.* Methylation of H3-Lysine 79 Is Mediated by a New Family of HMTases without a SET Domain. *Curr. Biol.* **12**, 1052–1058 (2002).

319. Van Leeuwen, F., Gafken, P. R. & Gottschling, D. E. Dot1p modulates silencing in yeast by methylation of the nucleosome core. *Cell* **109**, 745–756 (2002).

320. Lin, W.-J., Gary, J. D., Yang, M. C., Clarke, S. & Herschman, H. R. The Mammalian Immediate-early TIS21 Protein and the Leukemia-associated BTG1 Protein Interact with a Protein-arginine N-Methyltransferase *. *J. Biol. Chem.* **271**, 15034–15044 (1996).

321. Lewis, E. B. A gene complex controlling segmentation in Drosophila. *Nature* vol. 276 565–570 (1978).

322. Ingham, P. W. Trithorax: A new homoeotic mutation of Drosophila melanogaster - II. The role of trx+ after embryogenesis. *Wilhelm Roux's Arch. Dev. Biol.* **190**, 365–369 (1981).

323. Shearn, A. The ash-1, ash-2 and trithorax genes of Drosophila melanogaster are functionally related. *Genetics* **121**, 517–525 (1989).

324. Tsukada, Y. I. *et al.* Histone demethylation by a family of JmjC domain-containing proteins. *Nature* **439**, 811–816 (2006).

325. Hödl, M. & Basler, K. Transcription in the absence of histone H3.2 and H3K4 methylation. *Curr. Biol.* **22**, 2253–2257 (2012).

326. Pengelly, A. R., Copur, Ö., Jäckle, H., Herzig, A. & Müller, J. A histone mutant reproduces the phenotype caused by loss of histone-modifying factor polycomb. *Science (80-.).* **339**, 698–699 (2013).

327. Dorighi, K. M. *et al.* Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Mol. Cell* **66**, 568-576.e4 (2017).

328. Li, E. & Zhang, Y. DNA methylation in mammals. *Cold Spring Harb. Perspect. Biol.* **6**, (2014).

329. Johnson, T. B. & Coghill, R. D. Researches on pyrimidines. C111. The discovery of 5-methyl-cytosine in tuberculinic acid, the nucleic acid of the tubercle bacillus. *J. Am. Chem. Soc.* **47**, 2838–2844 (1925).

330. Hotchkiss, R. D. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *J. Biol. Chem.* **175**, 315–332 (1948).

331. Scarano, E. The control of gene function in cell differentiation and in embryogenesis. *Adv. Cytopharmacol.* **1**, 13–24 (1971).

332. Venner, H. & Reinert, H. Possible role of methylated DNA bases for the transcription of the genetic information. *Z. Allg. Mikrobiol.* **13**, 613–624 (1973).

333. Holliday, R. & Pugh, J. E. DNA Modification Mechanisms and Gene Activity During Development. *Science (80-.).* **187**, 226–232 (1975).

334. Riggs, A. D. X inactivation, differentiation, and DNA methylation. *Cytogenet. Genome Res.* **14**, 9–25 (1975).

335. Razin, A. & Riggs, A. D. DNA Methylation and gene function. *Science (80-.).* **210**, 604–610 (1980).

336. Bird, A. P. DNA methylation patterns and epigenetic memory. *Genes and Development* vol. 16 6–21 (2002).

337. Ehrlich, M. *et al.* Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucleic Acids Res.* **10**, 2709–2721 (1982).

338. Coulondre, C., Miller, J. H., Farabaugh, P. J. & Gilbert, W. Molecular basis of base substitution hotspots in Escherichia coli. *Nature* **274**, 775–780 (1978).

339. Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**, 1499–1504 (1980).

340. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

341. Deaton, A. M. & Bird, A. P. CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010–1022 (2011).

342. Schultz, M. D. *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**, 212–216 (2015).

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

344. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology* vol. 20 590–607 (2019).

345. Jones, P. A. Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* vol. 13 484–492 (2012).

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

347. Ball, M. P. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotechnol.* **27**, 361–368 (2009).

348. Weber, M. *et al.* Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* **39**, 457–466 (2007).

349. Galupa, R. & Heard, E. X-chromosome inactivation: A crossroads between chromosome architecture and gene regulation. *Annual Review of Genetics* vol. 52 535–566 (2018).

350. Hanna, C. W. & Kelsey, G. Features and mechanisms of canonical and noncanonical genomic imprinting. *Genes and Development* vol. 38 821–834 (2021).

351. Prendergast, G. C. & Ziff, E. B. Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region. *Science (80-.).* **251**, 186–189 (1991).

352. Gaston, K. & Fried, M. CpG methylation has differential effects on the binding of YY1 and ETS proteins to the bi-directional promoter of the Surf-1 and surf-2 genes. *Nucleic Acids Res.* **23**, 901–909 (1995).

353. Harrington, M. A., Jones, P. A., Imagawa, M. & Karin, M. Cytosine methylation does not affect binding of transcription factor Sp1. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 2066–2070 (1988).

354. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science (80-.).* **356**, (2017).

355. Zhu, H., Wang, G. & Qian, J. Transcription factors as readers and effectors of DNA methylation. *Nat. Rev. Genet.* **17**, 551–565 (2016).

356. Buck-Koehntop, B. A. & Defossez, P. A. On how mammalian transcription factors recognize methylated DNA. *Epigenetics* **8**, 131–137 (2013).

357. Baubec, T., Ivánek, R., Lienert, F. & Schübeler, D. Methylation-dependent and -independent genomic targeting principles of the mbd protein family. *Cell* **153**, 480–492 (2013).

358. Meehan, R. R., Lewis, J. D., McKay, S., Kleiner, E. L. & Bird, A. P. Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. *Cell* **58**, 499–507 (1989).

359. Filion, G. J. P. *et al.* A Family of Human Zinc Finger Proteins That Bind Methylated DNA and Repress Transcription. *Mol. Cell. Biol.* **26**, 169–181 (2006).

360. Unoki, M., Nishidate, T. & Nakamura, Y. ICBP90, an E2F-1 target, recruits HDAC1 and binds to methyl-CpG through its SRA domain. *Oncogene* **23**, 7601–7610 (2004).

361. Bostick, M. *et al.* UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science (80-.).* **317**, 1760–1764 (2007).

362. Du, Q., Luu, P. L., Stirzaker, C. & Clark, S. J. Methyl-CpG-binding domain proteins: Readers of the epigenome. *Epigenomics* vol. 7 1051–1073 (2015).

363. Felsenfeld, G. & Bell, A. C. Methylation of a CTCF-dependent boundary controls imprinted expressionof the *Igf2* gene. *Nature* **405**, 482–485 (2000).

364. Hark, A. T. *et al.* CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* **405**, 486–489 (2000).

365. Wiehle, L. *et al.* DNA (de)methylation in embryonic stem cells controls CTCF-dependent chromatin boundaries. *Genome Res.* **29**, 750–761 (2019).

366. Wang, H. *et al.* Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.* **22**, 1680–8 (2012).

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

368. Hashimoto, H. *et al.* Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA. *Mol. Cell* **66**, 711-720.e3 (2017).

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

370. Bourc'his, D., Xu, G. L., Lin, C. S., Bollman, B. & Bestor, T. H. Dnmt3L and the establishment of maternal genomic imprints. *Science (80-.).* **294**, 2536–2539 (2001).

371. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247–257 (1999).

372. Chen, T. & Li, E. Structure and Function of Eukaryotic DNA Methyltransferases. *Curr. Top. Dev. Biol.* **60**, 55–89 (2004).

373. Kato, Y. *et al.* Role of the Dnmt3 family in de novo methylation of imprinted and repetitive sequences during male germ cell development in the mouse. *Hum. Mol. Genet.* **16**, 2272–2280 (2007).

374. Gujar, H., Weisenberger, D. J. & Liang, G. The roles of human DNA methyltransferases and their isoforms in shaping the epigenome. *Genes* vol. 10 (2019).

375. Huntriss, J. *et al.* Expression of mRNAs for DNA Methyltransferases and Methyl-CpG-Binding Proteins in the Human Female Germ Line, Preimplantation Embryos, and Embryonic Stem Cells. *Mol. Reprod. Dev.* **67**, 323–336 (2004).

376. Shirane, K. *et al.* Mouse Oocyte Methylomes at Base Resolution Reveal Genome-Wide Accumulation of Non-CpG Methylation and Role of DNA Methyltransferases. *PLOS Genet.* **9**, e1003439 (2013).

377. Veland, N. *et al.* DNMT3L facilitates DNA methylation partly by maintaining DNMT3A stability in mouse embryonic stem cells. *Nucleic Acids Res.* **47**, 152–167 (2019).

378. Bestor, T. H. & Ingram, V. M. Two DNA methyltransferases from murine erythroleukemia cells: Purification, sequence specificity, and mode of interaction with DNA. *Proc. Natl. Acad. Sci. U. S. A.* **80**, 5559–5563 (1983).

379. Li, E., Bestor, T. H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**, 915–926 (1992).

380. Hervouet, E., Peixoto, P., Delage-Mourroux, R., Boyer-Guittaut, M. & Cartron, P. F. Specific or not specific recruitment of DNMTs for DNA methylation, an epigenetic dilemma. *Clinical Epigenetics* vol. 10 1–18 (2018).

381. Iida, T. *et al.* PCNA clamp facilitates action of DNA cytosine methyltransferase 1 on hemimethylated DNA. *Genes to Cells* **7**, 997–1007 (2002).

382. Chen, T., Ueda, Y., Dodge, J. E., Wang, Z. & Li, E. Establishment and Maintenance of Genomic Methylation Patterns in Mouse Embryonic Stem Cells by Dnmt3a and Dnmt3b. *Mol. Cell. Biol.* **23**, 5594–5605 (2003).

383. Liang, G. *et al.* Cooperativity between DNA Methyltransferases in the Maintenance Methylation of Repetitive Elements. *Mol. Cell. Biol.* **22**, 480–491 (2002).

384. Rhee, I. *et al.* DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature* **416**, 552–556 (2002).

385. Fatemi, M., Hermann, A., Gowher, H. & Jeltsch, A. Dnmt3a and Dnmt1 functionally cooperate during de novo methylation of DNA. *Eur. J. Biochem.* **269**, 4981–4984 (2002).

386. Hirasawa, R. *et al.* Maternal and zygotic Dnmt1 are necessary and sufficient for the maintenance of DNA methylation imprints during preimplantation development. *Genes Dev.* **22**, 1607 (2008).

387. Ito, S. *et al.* Role of tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**, 1129–1133 (2010).

388. Wu, H. & Zhang, Y. Reversing DNA methylation: Mechanisms, genomics, and biological functions. *Cell* vol. 156 45–68 (2014).

389. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science (80-.).* **324**, 930–935 (2009).

390. Wu, X. & Zhang, Y. TET-mediated active DNA demethylation: Mechanism, function and beyond. *Nature Reviews Genetics* vol. 18 517–534 (2017).

391. Ito, S. *et al.* Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science (80-.).* **333**, 1300–1303 (2011).

392. He, Y. F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science (80-.).* **333**, 1303–1307 (2011).

393. Maiti, A. & Drohat, A. C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: Potential implications for active demethylation of CpG sites. *J. Biol. Chem.* **286**, 35334–35338 (2011).

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

395. Melamed, P., Yosefzon, Y., David, C., Tsukerman, A. & Pnueli, L. Tet Enzymes, Variants, and Differential Effects on Function. *Front. Cell Dev. Biol.* **6**, (2018).

396. Dawlaty, M. M. *et al.* Combined Deficiency of Tet1 and Tet2 Causes Epigenetic Abnormalities but Is Compatible with Postnatal Development. *Dev. Cell* **24**, 310–323 (2013).

397. Dawlaty, M. M. *et al.* Tet1 is dispensable for maintaining pluripotency and its loss is compatible with embryonic and postnatal development. *Cell Stem Cell* **9**, 166–175 (2011).

398. Moran-Crusio, K. *et al.* Tet2 Loss Leads to Increased Hematopoietic Stem Cell Self-Renewal and Myeloid Transformation. *Cancer Cell* **20**, 11–24 (2011).

399. An, J. *et al.* Acute loss of TET function results in aggressive myeloid cancer in mice. *Nat. Commun.* **6**, 1–14 (2015).

400. Rasmussen, K. D. *et al.* TET2 binding to enhancers facilitates transcription factor recruitment in hematopoietic cells. *Genome Res.* **29**, 564–575 (2019).

401. Ziller, M. J. *et al.* Dissecting the Functional Consequences of De Novo DNA Methylation Dynamics in Human Motor Neuron Differentiation and Physiology. *Cell Stem Cell* **22**, 559-574.e9 (2018).

402. Challen, G. A. *et al.* Dnmt3a is essential for hematopoietic stem cell differentiation. *Nat. Genet.* **44**, 23–31 (2012).

403. Duymich, C. E., Charlet, J., Yang, X., Jones, P. A. & Liang, G. DNMT3B isoforms without catalytic activity stimulate gene body methylation as accessory proteins in somatic cells. *Nat. Commun.* **7**, 11453 (2016).

404. Barwick, B. G. *et al.* B cell activation and plasma cell differentiation are inhibited by de novo DNA methylation. *Nat. Commun. 2018 91* **9**, 1–14 (2018).

405. Ooi, S. K. T. *et al.* DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448**, 714–717 (2007).

406. Zhang, Y. *et al.* Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail. *Nucleic Acids Res.* **38**, 4246–4253 (2010).

407. Neri, F. *et al.* Dnmt3L antagonizes DNA methylation at bivalent promoters and favors DNA methylation at gene bodies in ESCs. *Cell* **155**, 121 (2013).

408. Dhayalan, A. *et al.* The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. *J. Biol. Chem.* **285**, 26114–26120 (2010).

409. Rothbart, S. B. *et al.* Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation. *Nat. Struct. Mol. Biol.* **19**, 1155–1160 (2012).

410. Viré, E. *et al.* The Polycomb group protein EZH2 directly controls DNA methylation. *Nature* **439**, 871–874 (2006).

411. Li, H. *et al.* The histone methyltransferase SETDB1 and the DNA methyltransferase DNMT3A interact directly and localize to promoters silenced in cancer cells. *J. Biol. Chem.* **281**, 19489–19500 (2006).

412. Fuks, F., Burgers, W. A., Brehm, A., Hughes-Davies, L. & Kouzarides, T. DNA methyltransferase Dnmt1 associates with histone deacetylase activity. *Nat. Genet.* **24**, 88–91 (2000).

413. Fuks, F., Burgers, W. A., Godin, N., Kasai, M. & Kouzarides, T. Dnmt3a binds deacetylases and is recruited by a sequence-specific repressor to silence transcription. *EMBO J.* **20**, 2536–2544 (2001).

414. Bird, A. P. CpG-Rich islands and the function of DNA methylation. *Nature* **321**, 209–213 (1986).

415. Gardiner-Garden, M. & Frommer, M. CpG Islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).

416. Illingworth, R. S. & Bird, A. P. CpG islands - 'A rough guide'. *FEBS Letters* vol. 583 1713–1720 (2009).

417. Irizarry, R. A. *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet. 2009 412* **41**, 178–186 (2009).

418. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288–295 (2011).

419. Sandoval, J. *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* **6**, 692–702 (2011).

420. Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics* **33**, 2381–2383 (2017).

421. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 1412–1417 (2006).

422. Tazi, J. & Bird, A. P. Alternative chromatin structure at CpG islands. *Cell* **60**, 909–920 (1990).

423. Gushchanskaya, E. S. *et al.* The clustering of CpG islands may constitute an important determinant of the 3D organization of interphase chromosomes. *Epigenetics* **9**, 951 (2014).

424. Ramirez-Carrozzi, V. R. *et al.* A Unifying Model for the Selective Regulation of Inducible Transcription by CpG Islands and Nucleosome Remodeling. *Cell* **138**, 114–128 (2009).

425. Mohn, F. *et al.* Lineage-Specific Polycomb Targets and De Novo DNA Methylation Define Restriction and Potential of Neuronal Progenitors. *Mol. Cell* **30**, 755–766 (2008).

426. Illingworth, R. S. *et al.* Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet.* **6**, e1001134 (2010).

427. Maunakea, A. K. *et al.* Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nat. 2010 4667303* **466**, 253–257 (2010).

428. Pachano, T. *et al.* Orphan CpG islands amplify poised enhancer regulatory activity and determine target gene responsiveness. *Nat. Genet.* **53**, 1036–1049 (2021).

429. Schübeler, D. Function and information content of DNA methylation. *Nature* vol. 517 321–326 (2015).

430. Williams, K. *et al.* TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343–349 (2011).

431. Wu, H. *et al.* Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature* **473**, 389–394 (2011).

432. Boulard, M., Edwards, J. R. & Bestor, T. H. FBXL10 protects Polycomb-bound genes from hypermethylation. *Nat. Genet.* **47**, 479–485 (2015).

433. Smallwood, S. A. *et al.* Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nat. Genet.* **43**, 811–814 (2011).

434. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet. 2013 143* **14**, 204–220 (2013).

435. Brenet, F. *et al.* DNA Methylation of the First Exon Is Tightly Linked to Transcriptional Silencing. *PLoS One* **6**, e14524 (2011).

436. Eckhardt, F. *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet. 2006 3812* **38**, 1378–1385 (2006).

437. Gal-Yam, E. N. *et al.* Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. *Proc. Natl. Acad. Sci.* **105**, 12979–12984 (2008).

438. Han, H. *et al.* DNA methylation directly silences genes with non-CpG island promoters and establishes a nucleosome occupied promoter. *Hum. Mol. Genet.* **20**, 4299 (2011).

439. Jones, P. A. & Baylin, S. B. The Epigenomics of Cancer. *Cell* vol. 128 683–692 (2007).

440. Toyota, M. *et al.* CpG island methylator phenotype in colorectal cancer. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 8681–8686 (1999).

441. Kane, M. F. *et al.* Methylation of the hMLH1 Promoter Correlates with Lack of Expression of hMLH1 in Sporadic Colon Tumors and Mismatch Repair-defective Human Tumor Cell Lines. *Cancer Res.* **57**, (1997).

442. Veigl, M. L. *et al.* Biallelic inactivation of hMLH1 by epigenetic gene silencing, a novel mechanism causing human MSI cancers. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 8698–8702 (1998).

443. Figueroa, M. E. *et al.* Genome-wide epigenetic analysis delineates a biologically distinct immature acute leukemia with myeloid/T-lymphoid features. *Blood* **113**, 2795–2804 (2009).

444. Gebhard, C. *et al.* Profiling of aberrant DNA methylation in acute myeloid leukemia reveals subclasses of CG-rich regions with epigenetic or genetic association. *Leukemia* **33**, 26–36 (2019).

445. Kelly, A. D. *et al.* A CpG island methylator phenotype in acute myeloid leukemia independent of IDH mutations and associated with a favorable outcome. *Leukemia* **31**, 2011–2019 (2017).

446. Schlesinger, Y. *et al.* Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat. Genet.* **39**, 232–236 (2007).

447. Ohm, J. E. *et al.* A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat. Genet.* **39**, 237–242 (2007).

448. Kumar, D., Cinghu, S., Oldfield, A. J., Yang, P. & Jothi, R. Decoding the function of bivalent chromatin in development and cancer. *Genome Res.* gr.275736.121 (2021) doi:10.1101/gr.275736.121.

449. Feltus, F. A., Lee, E. K., Costello, J. F., Plass, C. & Vertino, P. M. Predicting aberrant CpG island methylation. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 12253–12258 (2003).

450. Croce, L. Di *et al.* Methyltransferase Recruitment and DNA Hypermethylation of Target Promoters by an Oncogenic Transcription Factor. *Science (80-.).* **295**, 1079–1082 (2002).

451. Gama-Sosa, M. A. *et al.* The 5-methylcytosine content of DNA from human tumors. *Nucleic Acids Res.* **11**, 6883 (1983).

452. Feinberg, A. P. & Vogelstein, B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **301**, 89–92 (1983).

453. Weisenberger, D. J. *et al.* Analysis of repetitive element DNA methylation by MethyLight. *Nucleic Acids Res.* **33**, 6823–6836 (2005).

454. Pfeifer, G. P. & Rauch, T. A. DNA methylation patterns in lung carcinomas. *Seminars in Cancer Biology* vol. 19 181–187 (2009).

455. Carr, B. I., Reilly, J. G., Smith, S. S., Winberg, C. & Riggs, A. The tumorigenicity of 5-azacytidine in the male fischer rat. *Carcinogenesis* **5**, 1583–1590 (1984).

456. Yamada, Y. *et al.* Opposing effects of DNA hypomethylation on intestinal and liver carcinogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 13580–13585 (2005).

457. Chen, R. Z., Pettersson, U., Beard, C., Jackson-Grusby, L. & Jaenisch, R. DNA hypomethylation leads to elevated mutation rates. *Nature* **395**, 89–93 (1998).

458. Ehrlich, M. DNA hypomethylation in cancer cells. *Epigenomics* vol. 1 239–259 (2009).

459. Feinberg, A. P. & Vogelstein, B. Hypomethylation of ras oncogenes in primary human cancers. *Biochem. Biophys. Res. Commun.* **111**, 47–54 (1983).

460. Rainier, S. *et al.* Relaxation of imprinted genes in human cancer. *Nature* **362**, 747–749 (1993).

461. Suzuki, H., Ueda, R., Takahashi, T. & Takahashi, T. Altered imprinting in lung cancer. *Nat. Genet. 1994 64* **6**, 332–333 (1994).

462. Randhawa, G. S. *et al.* Loss of Imprinting in Disease Progression in Chronic Myelogenous Leukemia. *Blood* **91**, 3144–3147 (1998).

463. Holm, T. M. *et al.* Global loss of imprinting leads to widespread tumorigenesis in adult mice. *Cancer Cell* **8**, 275–285 (2005).

464. Ley, T. J. *et al.* DNMT3A mutations in acute myeloid leukemia. *N. Engl. J. Med.* **363**, 2424–33 (2010).

465. Yang, L., Rau, R. & Goodell, M. A. DNMT3A in haematological malignancies. *Nat. Rev. Cancer* **15**, 152–165 (2015).

466. Nosho, K. *et al.* DNMT3B expression might contribute to CpG island methylator phenotype in colorectal cancer. *Clin. Cancer Res.* **15**, 3663–3671 (2009).

467. Rajendran, G. *et al.* Epigenetic regulation of DNA methyltransferases: DNMT1 and DNMT3B in gliomas. *J. Neurooncol.* **104**, 483–494 (2011).

468. Gagliardi, M., Strazzullo, M. & Matarazzo, M. R. DNMT3B Functions: Novel Insights From Human Disease. *Front. Cell Dev. Biol.* **0**, 140 (2018).

469. Delhommeau, F. *et al.* Mutation in TET2 in Myeloid Cancers. *N. Engl. J. Med.* **360**, 2289–2301 (2009).

470. Abdel-Wahab, O. *et al.* Genetic characterization of TET1, TET2, and TET3 alterations in myeloid malignancies. *Blood* **114**, 144–147 (2009).

471. Fernandez, A. F. *et al.* A DNA methylation fingerprint of 1628 human samples. *Genome Res.* **22**, 407–419 (2012).

472. Bormann, F. *et al.* Cell-of-Origin DNA Methylation Signatures Are Maintained during Colorectal Carcinogenesis. *Cell Rep.* **23**, 3407–3418 (2018).

473. Zheng, C. & Xu, R. Predicting cancer origins with a DNA methylation-based deep neural network model. *PLoS One* **15**, e0226461 (2020).

474. Zhang, S. *et al.* Discriminating Origin Tissues of Tumor Cell Lines by Methylation Signatures and Dys-Methylated Rules. *Front. Bioeng. Biotechnol.* **8**, 507 (2020).

475. Miranda, T. B. & Jones, P. A. DNA methylation: The nuts and bolts of repression. *Journal of Cellular Physiology* vol. 213 384–390 (2007).

476. Laird, P. W. The power and the promise of DNA methylation markers. *Nature Reviews Cancer* vol. 3 253–266 (2003).

477. Meselson, M., Yuan, R. & Heywood, J. Restriction and modification of DNA. *Annu. Rev. Biochem.* **41**, 447–66 (1972).

478. Roizes, G. A Possible Structure for Calf Satellite DNA I. *Nucleic Acids Res.* **3**, 2677–2696 (1976).

479. Singer, J., Roberts-Ems, J. & Riggs, A. D. Methylation of mouse liver DNA studied by means of the restriction enzymes Msp I and Hpa II. *Science (80-.).* **203**, 1019–1021 (1979).

480. Cedar, H., Solage, A., Glaser, G. & Razin, A. Direct detection of methylated cytosine in DNA by use of the restriction enzyme Mspl. *Nucleic Acids Res.* **6**, 2125–2132 (1979).

481. Khulan, B. *et al.* Comparative isoschizomer profiling of cytosine methylation: The HELP assay. *Genome Res.* **16**, 1046–1055 (2006).

482. Harrison, A. & Parle-McDermott, A. DNA methylation: A timeline of methods and applications. *Frontiers in Genetics* vol. 2 74 (2011).

483. Hayatsu, H. Discovery of bisulfite-mediated cytosine conversion to uracil, the key reaction for DNA methylation analysis - A personal account. *Proceedings of the Japan Academy Series B: Physical and Biological Sciences* vol. 84 321–330 (2008).

484. Shapiro, R., Servis, R. E. & Wecher, M. Reactions of Uracil and Cytosine Derivatives with Sodium Bisulfite. A Specific Deamination Method. *J. Am. Chem. Soc.* **92**, 422–424 (1970).

485. Hayatsu, H., Wataya, Y. & Kai, K. The Addition of Sodium Bisulfite to Uracil and to Cytosine. *J. Am. Chem. Soc.* **92**, 724–726 (1970).

486. Wang, R. Y. H., Gehrke, C. W. & Ehrlich, M. Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues. *Nucleic Acids Res.* **8**, 4777–4790 (1980).

487. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5- methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 1827–1831 (1992).

488. Gitan, R. S., Shi, H., Chen, C. M., Yan, P. S. & Huang, T. H. M. Methylation-specific oligonucleotide microarray: A new potential for high-throughput methylation analysis. *Genome Res.* **12**, 158–164 (2002).

489. Meissner, A. *et al.* Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res.* **33**, 5868–5877 (2005).

490. Sano, H., Royer, H. D. & Sager, R. Identification of 5-methylcytosine in DNA fragments immobilized on nitrocellulose paper. *Proc. Natl. Acad. Sci. U. S. A.* **77**, 3581–3585 (1980).

491. Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.* **37**, 853–862 (2005).

492. Gebhard, C. *et al.* Genome-wide profiling of CpG methylation identifies novel targets of aberrant hypermethylation in myeloid leukemia. *Cancer Res.* **66**, 6118–6128 (2006).

493. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).

494. Lorthongpanich, C. *et al.* Single-cell DNA-methylation analysis reveals epigenetic chimerism in preimplantation embryos. *Science (80-.).* **341**, 1110–1112 (2013).

495. Smallwood, S. A. *et al.* Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014).

496. Farlik, M. *et al.* Single-Cell DNA Methylome Sequencing and Bioinformatic Inference of Epigenomic Cell-State Dynamics. *Cell Rep.* **10**, 1386–1397 (2015).

497. Farlik, M. *et al.* DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. *Cell Stem Cell* **19**, 808–822 (2016).

498. Jacob, F., Ullman, A. & Monod, J. [The promotor, a genetic element necessary to the expression of an operon]. *C. R. Hebd. Seances Acad. Sci.* **258**, 3125–318 (1964).

499. Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nature Reviews Genetics* vol. 21 71–87 (2020).

500. Carninci, P. *et al.* Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* **38**, 626–635 (2006).

501. Juven-Gershon, T. & Kadonaga, J. T. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Developmental Biology* vol. 339 225–229 (2010).

502. Thomas, M. C. & Chiang, C. M. The general transcription machinery and general cofactors. *Critical reviews in biochemistry and molecular biology* vol. 41 105–178 (2006).

503. Smale, S. T. & Baltimore, D. The 'initiator' as a transcription control element. *Cell* **57**, 103–113 (1989).

504. Tokusumi, Y., Ma, Y., Song, X., Jacobson, R. H. & Takada, S. The new core promoter element XCPE1 (X Core Promoter Element 1) directs activator-, mediator-, and TATA-binding protein-dependent but TFIID-independent RNA polymerase II transcription from TATA-less promoters. *Mol. Cell. Biol.* **27**, 1844–1858 (2007).

505. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. *Cell* **130**, 77–88 (2007).

506. Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–1320 (2014).

507. Scruggs, B. S. *et al.* Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Mol. Cell* **58**, 1101–1112 (2015).

508. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: From properties to genome-wide predictions. *Nature Reviews Genetics* vol. 15 272–286 (2014).

509. Moreau, P. *et al.* The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Res.* **9**, 6047–6068 (1981).

510. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).

511. Banerji, J., Olson, L. & Schaffner, W. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33**, 729–740 (1983).

512. Gillies, S. D., Morrison, S. L., Oi, V. T. & Tonegawa, S. A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell* **33**, 717–728 (1983).

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

514. Kim, T. K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).

515. Melgar, M. F., Collins, F. S. & Sethupathy, P. Discovery of active enhancers through bidirectional expression of short transcripts. *Genome Biol.* **12**, 1–11 (2011).

516. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).

517. Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C. & Cohen, B. A. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 19498–19503 (2012).

518. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science (80-.).* **339**, 1074–1077 (2013).

519. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* **24**, 1595–1602 (2014).

520. Sethi, A. *et al.* Supervised enhancer prediction with epigenetic pattern recognition and targeted validation. *Nat. Methods* **17**, 807–814 (2020).

521. Zhu, Y. *et al.* Predicting enhancer transcription and activity from chromatin modifications. *Nucleic Acids Res.* **41**, 10032–10043 (2013).

522. Kouno, T. *et al.* C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. *Nat. Commun.* **10**, 1–12 (2019).

523. van Arensbergen, J., van Steensel, B. & Bussemaker, H. J. In search of the determinants of enhancer-promoter interaction specificity. *Trends in Cell Biology* vol. 24 695–702 (2014).

524. Eychenne, T. *et al.* Functional interplay between Mediator and TFIIB in preinitiation complex assembly in relation to promoter architecture. *Genes Dev.* **30**, 2119–2132 (2016).

525. Boija, A. *et al.* CBP Regulates Recruitment and Release of Promoter-Proximal RNA Polymerase II. *Mol. Cell* **68**, 491-503.e5 (2017).

526. Moon, K. J. *et al.* The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. *Mol. Cell* **19**, 523–534 (2005).

527. Fukaya, T., Lim, B. & Levine, M. Enhancer Control of Transcriptional Bursting. *Cell* **166**, 358–368 (2016).

528. Koch, F. *et al.* Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat. Struct. Mol. Biol.* **18**, 956–963 (2011).

529. Kowalczyk, M. S. *et al.* Intragenic Enhancers Act as Alternative Promoters. *Mol. Cell* **45**, 447–458 (2012).

530. Dao, L. T. M. *et al.* Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat. Genet.* **49**, 1073–1081 (2017).

531. Andersson, R., Sandelin, A. & Danko, C. G. A unified architecture of transcriptional regulatory elements. *Trends Genet.* **31**, 426–433 (2015).

532. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).

533. Lovén, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320–334 (2013).

534. Pott, S. & Lieb, J. D. What are super-enhancers? *Nat. Genet.* **47**, 8–12 (2014).

535. Parker, S. C. J. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 17921–17926 (2013).

536. Grosveld, F., van Assendelft, G. B., Greaves, D. R. & Kollias, G. Position-independent, high-level expression of the human β-globin gene in transgenic mice. *Cell* **51**, 975–985 (1987).

537. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, (2013).

538. Mansour, M. R. *et al.* An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science (80-.).* **346**, 1373–1377 (2014).

539. Hnisz, D. *et al.* Convergence of Developmental and Oncogenic Signaling Pathways at Transcriptional Super-Enhancers. *Mol. Cell* **58**, 362–370 (2015).

540. Hay, D. *et al.* Genetic dissection of the α-globin super-enhancer in vivo. *Nat. Genet.* **48**, 895–903 (2016).

541. Shin, H. Y. *et al.* Hierarchy within the mammary STAT5-driven Wap super-enhancer. *Nat. Genet.* **48**, 904–911 (2016).

542. Huang, J. *et al.* Dissecting super-enhancer hierarchy based on chromatin interactions. *Nat. Commun.* **9**, 943 (2018).

543. Beagrie, R. A. *et al.* Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**, 519–524 (2017).

544. Sabari, B. R. *et al.* Coactivator condensation at super-enhancers links phase separation and gene control. *Science (80-.).* **361**, eaar3958 (2018).

545. Vos, E. S. M. *et al.* Interplay between CTCF boundaries and a super enhancer controls cohesin extrusion trajectories and gene expression. *Mol. Cell* **81**, 1–14 (2021).

546. Brand, A. H., Breeden, L., Abraham, J., Sternglanz, R. & Nasmyth, K. Characterization of a 'silencer' in yeast: A DNA sequence with properties opposite to those of a transcriptional enhancer. *Cell* **41**, 41–48 (1985).

547. Pang, B. & Snyder, M. P. Systematic identification of silencers in human cells. *Nat. Genet.* **52**, 254–263 (2020).

548. Sawada, S., Scarborough, J. D., Killeen, N. & Littman, D. R. A lineage-specific transcriptional silencer regulates CD4 gene expression during T lymphocyte development. *Cell* **77**, 917–929 (1994).

549. Huang, D., Petrykowska, H. M., Miller, B. F., Elnitski, L. & Ovcharenko, I. Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression. *Genome Res.* **29**, 657–667 (2019).

550. Cai, Y. *et al.* H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. *Nat. Commun.* **12**, 1–22 (2021).

551. Ngan, C. Y. *et al.* Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development. *Nat. Genet.* **52**, 264–272 (2020).

552. Gisselbrecht, S. S. *et al.* Transcriptional Silencers in Drosophila Serve a Dual Role as Transcriptional Enhancers in Alternate Cellular Contexts. *Mol. Cell* **77**, 324-337.e8 (2020).

553. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).

554. Soto, L. F. *et al.* Compendium of human transcription factor effector domains. *Mol. Cell* (2021) doi:10.1016/j.molcel.2021.11.007.

555. Latchman, D. S. *Transcription Factors: An Overview*. *Int. J. Biochem. Crll Biol* vol. 29 (1997).

556. Levati, E., Sartini, S., Ottonello, S. & Montanini, B. Dry and wet approaches for genome-wide functional annotation of conventional and unconventional transcriptional activators. *Computational and Structural Biotechnology Journal* vol. 14 262–270 (2016).

557. Lex, R. K. *et al.* GLI transcriptional repression regulates tissue-specific enhancer activity in response to hedgehog signaling. *Elife* **9**, (2020).

558. Brivanlou, A. H. & Darnell, J. E. Transcription: Signal transduction and the control of gene expression. *Science* vol. 295 813–818 (2002).

559. Ptashne, M. Specific binding of the λ phage repressor to λ DNA. *Nature* **214**, 232–234 (1967).

560. Gilbert, W. & Müller-Hill, B. The lac operator is DNA. *Proc. Natl. Acad. Sci. U. S. A.* **58**, 2415–2421 (1967).

561. Matsui, T., Segall, J., Weil, P. A. & Roeder, R. G. Multiple factors required for accurate initiation of transcription by purified RNA polymerase II. *J. Biol. Chem.* **255**, 11992–11996 (1980).

562. Dynan, W. S. & Tjian, R. The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. *Cell* **35**, 79–87 (1983).

563. Dynan, W. S. & Tjian, R. Control of eukaryotic messenger RNA synthesis by sequence-specific DNA-binding proteins. *Nature* vol. 316 774–778 (1985).

564. Johnson, P. F. & McKnight, S. L. Eukaryotic transcriptional regulatory proteins. *Annu. Rev. Biochem.* **58**, 799–839 (1989).

565. Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics* vol. 5 276–287 (2004).

566. Zambelli, F., Pesole, G. & Pavesi, G. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Brief. Bioinform.* **14**, 225–237 (2013).

567. Bailey, T. L. Discovering Novel Sequence Motifs with MEME. *Curr. Protoc. Bioinforma.* **00**, (2003).

568. Machanick, P. & Bailey, T. L. MEME-ChIP: Motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 (2011).

569. Lihu, A. & Holban, Ş. A review of ensemble methods for de novo motif discovery in ChIP-Seq data. *Brief. Bioinform.* **16**, 964–973 (2015).

570. Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**, D91-4 (2004).

571. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).

572. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

573. Mitchell, P. J. & Tjian, R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science (80-.).* **245**, 371–378 (1989).

574. Avellino, R. *et al.* An autonomous CEBPA enhancer specific for myeloid-lineage priming and neutrophilic differentiation. *Blood* **127**, 2991–3003 (2016).

575. Lin, Y. C. *et al.* A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. *Nat. Immunol.* **11**, 635–643 (2010).

576. Schmidt, F. *et al.* Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res.* **45**, 54–66 (2017).

577. Levo, M. & Segal, E. In pursuit of design principles of regulatory sequences. *Nature Reviews Genetics* vol. 15 453–468 (2014).

578. Brodsky, S. *et al.* Intrinsically Disordered Regions Direct Transcription Factor In Vivo Binding Specificity. *Mol. Cell* **79**, 459-471.e4 (2020).

579. Krois, A. S., Jane Dyson, H. & Wright, P. E. Long-range regulation of p53 DNA binding by its intrinsically disordered N-terminal transactivation domain. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E11302–E11310 (2018).

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

581. Mullen, A. C. *et al.* Master transcription factors determine cell-type-specific responses to TGF-β signaling. *Cell* **147**, 565–576 (2011).

582. Morgunova, E. & Taipale, J. Structural perspective of cooperative transcription factor binding. *Current Opinion in Structural Biology* vol. 47 1–8 (2017).

583. Johnson, A. D., Meyer, B. J. & Ptashne, M. Interactions between DNA-bound repressors govern regulation by the λ phage repressor. *Proc. Natl. Acad. Sci. U. S. A.* **76**, 5061–5065 (1979).

584. Reiter, F., Wienerroither, S. & Stark, A. Combinatorial function of transcription factors and cofactors. *Current Opinion in Genetics and Development* vol. 43 73–81 (2017).

585. Frietze, S. & Farnham, P. J. Transcription factor effector domains. *Subcell. Biochem.* **52**, 261–277 (2011).

586. Kim, S. Il, Bresnick, E. H. & Bultman, S. J. BRG1 directly regulates nucleosome structure and chromatin looping of the α globin locus to activate transcription. *Nucleic Acids Res.* **37**, 6019–6027 (2009).

587. Eberhardy, S. R. & Farnham, P. J. Myc Recruits P-TEFb to Mediate the Final Step in the Transcriptional Activation of the cad Promoter. *J. Biol. Chem.* **277**, 40156–40162 (2002).

588. Stadhouders, R., Filion, G. J. & Graf, T. Transcription factors and 3D genome conformation in cell-fate decisions. *Nature* vol. 569 345–354 (2019).

589. Novershtern, N. *et al.* Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144**, 296–309 (2011).

590. Laurenti, E. *et al.* The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment. *Nat. Immunol.* **14**, 756–763 (2013).

591. Clien, L. *et al.* Transcriptional diversity during lineage commitment of human blood progenitors. *Science (80-.).* **345**, (2014).

592. Tsai, F. Y. *et al.* An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature* **371**, 221–226 (1994).

593. Okuda, T., Van Deursen, J., Hiebert, S. W., Grosveld, G. & Downing, J. R. AML1, the target of multiple chromosomal translocations in human leukemia, is essential for normal fetal liver hematopoiesis. *Cell* **84**, 321–330 (1996).

594. Porcher, C. *et al.* The T cell leukemia oncoprotein SCL/tal-1 is essential for development of all hematopoietic lineages. *Cell* **86**, 47–57 (1996).

595. Laslo, P. *et al.* Multilineage Transcriptional Priming and Determination of Alternate Hematopoietic Cell Fates. *Cell* **126**, 755–766 (2006).

596. Scott, E. W., Simon, M. C., Anastasi, J. & Singh, H. Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages. *Science (80-.).* **265**, 1573–1577 (1994).

597. McKercher, S. R. *et al.* Targeted disruption of the PU.1 gene results in multiple hematopoietic abnormalities. *EMBO J.* **15**, 5647 (1996).

598. Nerlov, C. & Graf, T. PU.1 induces myeloid lineage commitment in multipotent hematopoietic progenitors. *Genes Dev.* **12**, 2403–2412 (1998).

599. Voso, M. T. *et al.* Inhibition of hematopoiesis by competitive binding of transcription factor PU.1. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 7932–7936 (1994).

600. DeKoter, R. P. & Singh, H. Regulation of B lymphocyte and macrophage development by graded expression of PU.1. *Science* **288**, 1439–1442 (2000).

601. Anderson, M. K., Weiss, A. H., Hernandez-Hoyos, G., Dionne, C. J. & Rothenberg, E. V. Constitutive expression of PU.1 in fetal hematopoietic progenitors blocks T cell development at the pro-T cell stage. *Immunity* **16**, 285–296 (2002).

602. Laiosa, C. V., Stadtfeld, M., Xie, H., de Andres-Aguayo, L. & Graf, T. Reprogramming of Committed T Cell Progenitors to Macrophages and Dendritic Cells by C/EBPα and PU.1 Transcription Factors. *Immunity* **25**, 731–744 (2006).

603. Del Real, M. M. & Rothenberg, E. V. Architecture of a lymphomyeloid developmental switch controlled by PU.1, notch and Gata3. *Development* **140**, 1207–1219 (2013).

604. Scott, L. M., Civin, C. I., Rorth, P. & Friedman, A. D. A novel temporal expression pattern of three C/EBP family members in differentiating myelomonocytic cells. *Blood* **80**, 1725–1735 (1992).

605. Zhang, D. E. *et al.* Absence of granulocyte colony-stimulating factor signaling and neutrophil development in CCAAT enhancer binding protein α-deficient mice. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 569–574 (1997).

606. Radomska, H. S. *et al.* CCAAT/enhancer binding protein alpha is a regulatory switch sufficient for induction of granulocytic development from bipotential myeloid progenitors. *Mol. Cell. Biol.* **18**, 4301–4314 (1998).

607. Ma, O., Hong, S. H., Guo, H., Ghiaur, G. & Friedman, A. D. Granulopoiesis requires increased C/EBPα compared to monopoiesis, correlated with elevated Cebpa in immature G-CSF receptor versus M-CSF receptor expressing cells. *PLoS One* **9**, (2014).

608. Wang, D., D'Costa, J., Civin, C. I. & Friedman, A. D. C/EBPalpha directs monocytic commitment of primary myeloid progenitors. *Blood* **108**, 1223–1229 (2006).

609. Kulessa, H., Frampton, J. & Graf, T. GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts. *Genes Dev.* **9**, 1250–1262 (1995).

610. Visvader, J. E., Elefanty, A. G., Strasser, A. & Adams, J. M. GATA-1 but not SCL induces megakaryocytic differentiation in an early myeloid line. *EMBO J.* **11**, 4557–4564 (1992).

611. Visvader, J. E., Crossley, M., Hill, J., Orkin, S. H. & Adams, J. M. The C-terminal zinc finger of GATA-1 or GATA-2 is sufficient to induce megakaryocytic differentiation of an early myeloid cell line. *Mol. Cell. Biol.* **15**, 634–641 (1995).

612. Zhang, P. *et al.* Negative cross-talk between hematopoietic regulators: GATA proteins repress PU.1. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 8705–8710 (1999).

613. Reddy, V. A. *et al.* Granulocyte inducer C/EBPα inactivates the myeloid master regulator PU.1: Possible role in lineage commitment decisions. *Blood* **100**, 483–490 (2002).

614. Dahl, R., Iyer, S. R., Owens, K. S., Cuylear, D. D. & Simon, M. C. The transcriptional repressor GFI-1 antagonizes PU.1 activity through protein-protein interaction. *J. Biol. Chem.* **282**, 6473–6483 (2007).

615. Usui, T. *et al.* T-bet regulates Th1 responses through essential effects on GATA-3 function rather than on IFNG gene acetylation and transcription. *J. Exp. Med.* **203**, 755–66 (2006).

616. Hohaus, S. *et al.* PU.1 (Spi-1) and C/EBP alpha regulate expression of the granulocyte-macrophage colony-stimulating factor receptor alpha gene. *Mol. Cell. Biol.* **15**, 5830–5845 (1995).

617. Walsh, J. C. *et al.* Cooperative and antagonistic interplay between PU.1 and GATA-2 in the specification of myeloid cell fates. *Immunity* **17**, 665–676 (2002).

618. Wilson, N. K. *et al.* Combinatorial transcriptional control in blood stem/progenitor cells: Genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* **7**, 532–544 (2010).

619. Pimanda, J. E. *et al.* Gata2, Fli1, and Scl form a recursively wired gene-regulatory circuit during early hematopoietic development. *Proc. Natl. Acad. Sci.* **104**, 17692–17697 (2007).

620. Hamey, F. K. *et al.* Reconstructing blood stem cell regulatory network models from single-cell molecular profiles. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 5822–5829 (2017).

621. Iwasaki, H. *et al.* The order of expression of transcription factors directs hierarchical specification of hematopoietic lineages. *Genes Dev.* **20**, 3010–3021 (2006).

622. Dahl, R. *et al.* Regulation of macrophage and neutrophil cell fates by the PU.1:C/EBPalpha ratio and granulocyte colony-stimulating factor. *Nat. Immunol.* **4**, 1029–1036 (2003).

623. Hu, M. *et al.* Multilineage gene expression precedes commitment in the hemopoietic system. *Genes Dev.* **11**, 774–785 (1997).

624. Rieger, M. A., Hoppe, P. S., Smejkal, B. M., Eitelhuber, A. C. & Schroeder, T. Hematopoietic cytokines can instruct lineage choice. *Science (80-.).* **325**, 217–218 (2009).

625. Grover, A. *et al.* Erythropoietin guides multipotent hematopoietic progenitor cells toward an erythroid fate. *J. Exp. Med.* **211**, 181 (2014).

626. Mossadegh-Keller, N. *et al.* M-CSF instructs myeloid lineage fate in single haematopoietic stem cells. *Nature* **497**, 239–243 (2013).

627. Petruk, S. *et al.* Structure of Nascent Chromatin Is Essential for Hematopoietic Lineage Specification. *Cell Rep.* **19**, 295–306 (2017).

628. Kloetgen, A., Thandapani, P., Tsirigos, A. & Aifantis, I. 3D Chromosomal Landscapes in Hematopoiesis and Immunity. *Trends in Immunology* vol. 40 809–824 (2019).

629. Chen, C. *et al.* Spatial Genome Re-organization between Fetal and Adult Hematopoietic Stem Cells. *Cell Rep.* **29**, 4200-4211.e7 (2019).

630. Zhang, C. *et al.* tagHi-C Reveals 3D Chromatin Architecture Dynamics during Mouse Hematopoiesis. *Cell Rep.* **32**, 108206 (2020).

631. Viny, A. D. *et al.* Cohesin Members Stag1 and Stag2 Display Distinct Roles in Chromatin Accessibility and Topological Control of HSC Self-Renewal and Differentiation. *Cell Stem Cell* (2019) doi:10.1016/J.STEM.2019.08.003.

632. Hu, G. *et al.* Transformation of Accessible Chromatin and 3D Nucleome Underlies Lineage Commitment of Early T Cells. *Immunity* **48**, 227-242.e8 (2018).

633. Johanson, T. M. *et al.* Transcription-factor-mediated supervision of global genome architecture maintains B cell identity. *Nat. Immunol.* **19**, 1257–1264 (2018).

634. Ji, H. *et al.* Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature* **467**, 338–342 (2010).

635. Bock, C. *et al.* DNA Methylation Dynamics during In Vivo Differentiation of Blood and Skin Stem Cells. *Mol. Cell* **47**, 633–647 (2012).

636. Figueroa, M. E. *et al.* Leukemic IDH1 and IDH2 Mutations Result in a Hypermethylation Phenotype, Disrupt TET2 Function, and Impair Hematopoietic Differentiation. *Cancer Cell* **18**, 553–567 (2010).

637. Cimmino, L. *et al.* TET1 is a tumor suppressor of hematopoietic malignancy. *Nat. Immunol.* **16**, 653–662 (2015).

638. Ono, R. *et al.* Tet1 is not required for myeloid leukemogenesis by MLL-ENL in novel mouse models. *PLoS One* **16**, e0248425 (2021).

639. Orlanski, S. *et al.* Tissue-specific DNA demethylation is required for proper B-cell differentiation and function. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 5018–5023 (2016).

640. Tadokoro, Y., Ema, H., Okano, M., Li, E. & Nakauchi, H. De novo DNA methyltransferase is essential for self-renewal, but not for differentiation, in hematopoietic stem cells. *J. Exp. Med.* **204**, 715–722 (2007).

641. Challen, G. A. *et al.* Dnmt3a and Dnmt3b Have Overlapping and Distinct Functions in Hematopoietic Stem Cells. *Cell Stem Cell* **15**, 350–364 (2014).

642. Bröske, A. M. *et al.* DNA methylation protects hematopoietic stem cell multipotency from myeloerythroid restriction. *Nat. Genet.* **41**, 1207–1215 (2009).

643. Trowbridge, J. J., Snow, J. W., Kim, J. & Orkin, S. H. DNA Methyltransferase 1 Is Essential for and Uniquely Regulates Hematopoietic Stem and Progenitor Cells. *Cell Stem Cell* **5**, 442–449 (2009).

644. Cole, C. B. *et al.* Haploinsufficiency for DNA methyltransferase 3A predisposes hematopoietic cells to myeloid malignancies. *J. Clin. Invest.* **127**, 3657–3674 (2017).

645. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).

646. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e16 (2018).

647. Martin, E. W. *et al.* Chromatin accessibility maps provide evidence of multilineage gene priming in hematopoietic stem cells. *Epigenetics and Chromatin* **14**, 1–15 (2021).

648. Ranzoni, A. M. *et al.* Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis. *Cell Stem Cell* **28**, 472-487.e7 (2021).

649. Han, L. *et al.* Chromatin remodeling mediated by ARID1A is indispensable for normal hematopoiesis in mice. *Leukemia* **33**, 2291–2305 (2019).

650. Vradii, D. *et al.* Brg1, the ATPase subunit of the SWI/SNF chromatin remodeling complex, is required for myeloid differentiation to granulocytes. *J. Cell. Physiol.* **206**, 112–118 (2006).

651. Witzel, M. *et al.* Chromatin-remodeling factor SMARCD2 regulates transcriptional networks controlling differentiation of neutrophil granulocytes. *Nat. Genet.* **49**, 742–752 (2017).

652. Liu, L. *et al.* The chromatin remodeling subunit Baf200 promotes normal hematopoiesis and inhibits leukemogenesis. *J. Hematol. Oncol.* **11**, 1–16 (2018).

653. Bakshi, R. *et al.* The human SWI/SNF complex associates with RUNX1 to control transcription of hematopoietic target genes. *J. Cell. Physiol.* **225**, 569–576 (2010).

654. Lara-Astiaso, D. *et al.* Immunogenetics. Chromatin state dynamics during blood formation. *Science* **345**, 943–9 (2014).

655. Cui, K. *et al.* Chromatin Signatures in Multipotent Human Hematopoietic Stem Cells Indicate the Fate of Bivalent Genes during Differentiation. *Cell Stem Cell* **4**, 80–93 (2009).

656. Wei, G. *et al.* Global Mapping of H3K4me3 and H3K27me3 Reveals Specificity and Plasticity in Lineage Fate Determination of Differentiating CD4+ T Cells. *Immunity* **30**, 155–167 (2009).

657. Mochizuki-Kashio, M. *et al.* Dependency on the polycomb gene Ezh2 distinguishes fetal from adult hematopoietic stem cells. *Blood* **118**, 6553–6561 (2011).

658. Su, I. H. *et al.* Ezh2 controls B cell development through histone H3 methylation and Igh rearrangement. *Nat. Immunol.* **4**, 124–131 (2003).

659. Yin, J. *et al.* Ezh2 regulates differentiation and function of natural killer cells through histone methyltransferase activity. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 15988–15993 (2015).

660. Hidalgo, I. *et al.* Ezh1 is required for hematopoietic stem cell maintenance and prevents senescence-like cell cycle arrest. *Cell Stem Cell* **11**, 649–662 (2012).

661. Xie, H. *et al.* Polycomb repressive complex 2 regulates normal hematopoietic stem cell function in a developmental-stage-specific manner. *Cell Stem Cell* **14**, 68–80 (2014).

662. Majewski, I. J. *et al.* Opposing roles of polycomb repressive complexes in hematopoietic stem and progenitor cells. *Blood* **116**, 731–739 (2010).

663. Kamminga, L. M. *et al.* The Polycomb group gene Ezh2 prevents hematopoietic stem cell exhaustion. *Blood* **107**, 2170–2179 (2006).

664. Zheng, L. *et al.* Utx loss causes myeloid transformation. *Leukemia* **32**, 1458–1465 (2018).

665. Kerenyi, M. A. *et al.* Histone demethylase Lsd1 represses hematopoietic stem and progenitor cell signatures during blood cell maturation. *Elife* **2**, (2013).

666. Saleque, S., Kim, J., Rooke, H. M. & Orkin, S. H. Epigenetic regulation of hematopoietic differentiation by Gfi-1 and Gfi-1b is mediated by the cofactors CoREST and LSD1. *Mol. Cell* **27**, 562–572 (2007).

667. Rebel, V. I. *et al.* Distinct roles for CREB-binding protein and p300 in hematopoietic stem cell self-renewal. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 14789–14794 (2002).

668. Kung, A. L. *et al.* Gene dose-dependent control of hematopoiesis and hematologic tumor suppression by CBP. *Genes Dev.* **14**, 272–277 (2000).

669. Oike, Y. *et al.* Mice Homozygous for a Truncated Form of CREB-Binding Protein Exhibit Defects in Hematopoiesis and Vasculo-angiogenesis. *Blood* **93**, 2771–2779 (1999).

670. Bedford, D. C., Kasper, L. H., Fukuyama, T. & Brindle, P. K. Target gene context influences the transcriptional requirement for the KAT3 family of CBP and p300 histone acetyltransferases. *Epigenetics* vol. 5 9–15 (2010).

671. Dancy, B. M. & Cole, P. A. Protein lysine acetylation by p300/CBP. *Chem. Rev.* **115**, 2419–2452 (2015).

672. Henry, R. A., Kuo, Y. M. & Andrews, A. J. Differences in specificity and selectivity between CBP and p300 acetylation of histone H3 and H3/H4. *Biochemistry* **52**, 5746–5759 (2013).

673. Chan, W.-I. *et al.* The Transcriptional Coactivator Cbp Regulates Self-Renewal and Differentiation in Adult Hematopoietic Stem Cells. *Mol. Cell. Biol.* **31**, 5046–5060 (2011).

674. Kasper, L. H. *et al.* Conditional Knockout Mice Reveal Distinct Functions for the Global Transcriptional Coactivators CBP and p300 in T-Cell Development. *Mol. Cell. Biol.* **26**, 789–809 (2006).

675. Xu, W. *et al.* Global transcriptional coactivators CREB-binding protein and p300 are highly essential collectively but not individually in peripheral B cells. *Blood* **107**, 4407–4416 (2006).

676. Wang, P., Wang, Z. & Liu, J. Role of HDACs in normal and malignant hematopoiesis. *Molecular Cancer* vol. 19 1–21 (2020).

677. Heideman, M. R. *et al.* Sin3a-associated Hdac1 and Hdac2 are essential for hematopoietic stem cell homeostasis and contribute differentially to hematopoiesis. *Haematologica* **99**, 1292–1303 (2014).

678. Wada, T. *et al.* Expression levels of histone deacetylases determine the cell fate of hematopoietic progenitors. *J. Biol. Chem.* **284**, 30673–30683 (2009).

679. Summers, A. R. *et al.* HDAC3 is essential for DNA replication in hematopoietic progenitor cells. *J. Clin. Invest.* **123**, 3112–3123 (2013).

680. Wang, H. *et al.* SIRT6 Controls Hematopoietic Stem Cell Homeostasis through Epigenetic Regulation of Wnt Signaling. *Cell Stem Cell* **18**, 495–507 (2016).

681. Hua, W. K. *et al.* HDAC8 regulates long-term hematopoietic stem-cell maintenance under stress by modulating p53 activity. *Blood* **130**, 2619 (2017).

682. Rimmelé, P. *et al.* Aging-like phenotype and defective lineage specification in SIRT1-deleted hematopoietic stem and progenitor cells. *Stem Cell Reports* **3**, 44–59 (2014).

683. Short, N. J., Rytting, M. E. & Cortes, J. E. Acute myeloid leukaemia. *Lancet (London, England)* **392**, 593–606 (2018).

684. Yamashita, M., Dellorusso, P. V., Olson, O. C. & Passegué, E. Dysregulated haematopoietic stem cell behaviour in myeloid leukaemogenesis. *Nature Reviews Cancer* vol. 20 365–382 (2020).

685. Krivtsov, A. V. *et al.* Transformation from committed progenitor to leukaemia stem cell initiated by MLL-AF9. *Nature* **442**, 818–822 (2006).

686. Goardon, N. *et al.* Coexistence of LMPP-like and GMP-like leukemia stem cells in acute myeloid leukemia. *Cancer Cell* **19**, 138–152 (2011).

687. Döhner, H. *et al.* Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* vol. 129 424–447 (2017).

688. Bennett, J. M. *et al.* Proposals for the Classification of the Acute Leukaemias French-American-British (FAB) Co-operative Group. *Br. J. Haematol.* **33**, 451–458 (1976).

689. Bennett, J. M. *et al.* Criteria for the diagnosis of acute leukemia of megakaryocyte lineage (M7). A report of the French-American-British Cooperative Group. *Ann. Intern. Med.* **103**, 460–462 (1985).

690. Bennett, J. M. *et al.* Proposed revised criteria for the classification of acute myeloid leukemia. A report of the French-American-British Cooperative Group. *Ann. Intern. Med.* **103**, 620–625 (1985).

691. Vardiman, J. W., Harris, N. L. & Brunning, R. D. The World Health Organization (WHO) classification of the myeloid neoplasms. *Blood* **100**, 2292–302 (2002).

692. Khoury, J. D. *et al.* The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Myeloid and Histiocytic/Dendritic Neoplasms. *Leukemia* **36**, 1703–1719 (2022).

693. Arber, D. A. *et al.* International Consensus Classification of Myeloid Neoplasms and Acute Leukemias: integrating morphologic, clinical, and genomic data. *Blood* **140**, 1200–1228 (2022).

694. Lindsley, R. C. *et al.* Acute myeloid leukemia ontogeny is defined by distinct somatic mutations. *Blood* **125**, 1367–1376 (2015).

695. Østgård, L. S. G. *et al.* Epidemiology and clinical significance of secondary and therapy-related acute myeloid leukemia: A national population-based cohort study. *J. Clin. Oncol.* **33**, 3641–3649 (2015).

696. Hulegårdh, E. *et al.* Characterization and prognostic features of secondary acute myeloid leukemia in a population-based setting: A report from the Swedish Acute Leukemia Registry. *Am. J. Hematol.* **90**, 208–214 (2015).

697. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer statistics, 2022. *CA. Cancer J. Clin.* **72**, 7–33 (2022).

698. Dong, Y. *et al.* Leukemia incidence trends at the global, regional, and national level between 1990 and 2017. *Exp. Hematol. Oncol.* **9**, 1–11 (2020).

699. Fitzmaurice, C. *et al.* Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study. *JAMA Oncol.* **3**, 524 (2017).

700. Noone, A. *et al.* Acute Myeloid Leukemia - Cancer Stat Facts. *Https://Seer.Cancer.Gov/* 1–3 (2018).

701. Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–278 (2012).

702. Döhner, H., Wei, A. H. & Löwenberg, B. Towards precision medicine for AML. *Nature Reviews Clinical Oncology* vol. 18 577–590 (2021).

703. Coombs, C. C., Tallman, M. S. & Levine, R. L. Molecular therapy for acute myeloid leukaemia. *Nat. Rev. Clin. Oncol.* **13**, 305–318 (2016).

704. Grove, C. S. & Vassiliou, G. S. Acute myeloid leukaemia: A paradigm for the clonal evolution of cancer? *DMM Disease Models and Mechanisms* vol. 7 941–951 (2014).

705. Nowell, P. C. The clonal evolution of tumor cell populations. *Science (80-.).* **194**, 23–28 (1976).

706. McCulloch, E. A., Buick, R. N., Lan, S. & Till, J. E. Differentiation in human myeloblastic leukemia studied in cell culture. *American Journal of Pathology* vol. 89 449–458 (1977).

707. Walter, M. J. *et al.* Clonal Architecture of Secondary Acute Myeloid Leukemia. *N. Engl. J. Med.* **366**, 1090–1098 (2012).

708. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506–510 (2012).

709. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).

710. Doolittle, W. F. & Brunet, T. D. P. On causal roles and selected effects: Our genome is mostly junk. *BMC Biol.* **15**, 1–9 (2017).

711. Ohno, S. So much 'junk' DNA in our genome. *Brookhaven Symp. Biol.* **23**, 366–70 (1972).

712. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041. e21 (2017).

713. Ostrow, S. L., Barshir, R., DeGregori, J., Yeger-Lotem, E. & Hershberg, R. Cancer Evolution Is Associated with Pervasive Positive Selection on Globally Expressed Genes. *PLOS Genet.* **10**, e1004239 (2014).

714. Fialkow, P. J., Jacobson, R. J. & Papayannopoulou, T. Chronic myelocytic leukemia: Clonal origin in a stem cell common to the granulocyte, erythrocyte, platelet and monocyte/macrophage. *Am. J. Med.* **63**, 125–130 (1977).

715. Fialkow, P. J., Gartler, S. M. & Yoshida, A. Clonal origin of chronic myelocytic leukemia in man. *Proc. Natl. Acad. Sci. U. S. A.* **58**, 1468–1471 (1967).

716. Moore, M. A. S., Williams, N. & Metcalf, D. In vitro colony formation by normal and leukemic human hematopoietic cells: Characterization of the colony-forming cells. *J. Natl. Cancer Inst.* **50**, 603–623 (1973).

717. Duttera, M. J., Whang-Peng, J., Bull, J. M. C. & Carbone, P. P. Cytogenetically abnormal cells in vitro in acute leukaemia. *Lancet (London, England)* **1**, 715–8 (1972).

718. Blackstock, A. M. & Garson, O. M. Direct evidence for involvement of erythroid cells in acute myeloblastic leukaemia. *Lancet (London, England)* **2**, 1178–9 (1974).

719. Minden, M. D., Till, J. E. & McCulloch, E. A. Proliferative state of blast cell progenitors in acute myelobastic leukemia (AML). *Blood* **52**, 592–600 (1978).

720. Buick, R. N., Minden, M. D. & McCulloch, E. A. Self-renewal in culture of proliferative blast progenitor cells in acute myeloblastic leukemia. *Blood* **54**, 95–104 (1979).

721. Griffin, J. & Lowenberg, B. Clonogenic cells in acute myeloblastic leukemia. *Blood* **68**, 1185–1195 (1986).

722. Lapidot, T. *et al.* A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367**, 645–648 (1994).

723. Haase, D. *et al.* Evidence for malignant transformation in acute myeloid leukemia at the level of early hematopoietic stem cells by cytogenetic analysis of CD34+ subpopulations. *Blood* **86**, 2906–2912 (1995).

724. Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat. Med.* **3**, 730–737 (1997).

725. Thomas, D. & Majeti, R. Biology and relevance of human acute myeloid leukemia stem cells. *Blood* vol. 129 1577–1585 (2017).

726. Horton, S. J. & Huntly, B. J. P. Recent advances in acute myeloid leukemia stem cell biology. *Haematologica* vol. 97 966–974 (2012).

727. Bloomfield, C. D. Acute non-lymphocytic leukemia. *Minn. Med.* **62**, 529–531 (1979).

728. Fialkow, P. *et al.* Acute nonlymphocytic leukemia: heterogeneity of stem cell origin. *Blood* **57**, 1068–1073 (1981).

729. Turhan, A. G. *et al.* Highly Purified Primitive Hematopoietic Stem Cells Are PML-RARA Negative and Generate Nonclonal Progenitors in Acute Promyelocytic Leukemia. *Blood* **85**, 2154–2161 (1995).

730. Cozzio, A. *et al.* Similar MLL-associated leukemias arising from self-renewing stem cells and short-lived myeloid progenitors. *Genes Dev.* **17**, 3029–3035 (2003).

731. Huntly, B. J. P. *et al.* MOZ-TIF2, but not BCR-ABL, confers properties of leukemic stem cells to committed murine hematopoietic progenitors. *Cancer Cell* **6**, 587–596 (2004).

732. Wang, Y. *et al.* The wnt/β-catenin pathway is required for the development of leukemia stem cells in AML. *Science (80-.).* **327**, 1650–1653 (2010).

733. Krivtsov, A. V. *et al.* Cell of origin determines clinically relevant subtypes of MLL-rearranged AML. *Leukemia* **27**, 852–860 (2013).

734. Chopra, M. & Bohlander, S. K. The cell of origin and the leukemia stem cell in acute myeloid leukemia. *Genes Chromosom. Cancer* **58**, 850–858 (2019).

735. Miyamoto, T., Weissman, I. L. & Akashi, K. AML1/ETO-expressing nonleukemic stem cells in acute myelogenous leukemia with 8;21 chromosomal translocation. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 7521–7526 (2000).

736. Corces-Zimmerman, M. R., Hong, W. J., Weissman, I. L., Medeiros, B. C. & Majeti, R. Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 2548–2553 (2014).

737. Jan, M. *et al.* Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Sci. Transl. Med.* **4**, 149ra118 (2012).

738. Shlush, L. I. *et al.* Identification of pre-leukaemic haematopoietic stem cells in acute leukaemia. *Nature* **506**, 328–333 (2014).

739. Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).

740. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).

741. Genovese, G. *et al.* Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).

742. Abelson, S. *et al.* Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400–404 (2018).

743. McKenzie, M. D. *et al.* Interconversion between Tumorigenic and Differentiated States in Acute Myeloid Leukemia. *Cell Stem Cell* **25**, 258-272.e9 (2019).

744. Jamieson, C. H. M. *et al.* Granulocyte–Macrophage Progenitors as Candidate Leukemic Stem Cells in Blast-Crisis CML. *N. Engl. J. Med.* **351**, 657–667 (2004).

745. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).

746. Walter, M. J. *et al.* Acquired copy number alterations in adult acute myeloid leukemia genomes. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 12950–12955 (2009).

747. Jaiswal, S. & Ebert, B. L. Clonal hematopoiesis in human aging and disease. *Science* vol. 366 (2019).

748. Deguchi, K. & Gilliland, D. G. Cooperativity between mutations in tyrosine kinases and in hematopoietic transcription factors in AML. *Leukemia* **16**, 740–744 (2002).

749. Kelly, L. M. *et al.* PML/RARα and FLT3-ITD induce an APL-like disease in a mouse model. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 8283–8288 (2002).

750. Groffen, J. *et al.* Philadelphia chromosomal breakpoints are clustered within a limited region, bcr, on chromosome 22. *Cell* **36**, 93–99 (1984).

751. Nowell, P. C. & Hungerford, D. A minute chromosome in human chronic granulocytic leukemia. *Science (80-.).* **132**, (1960).

752. Rowley, J. D. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243**, 290–293 (1973).

753. Zhang, X. & Ren, R. Bcr-Abl Efficiently Induces a Myeloproliferative Disease and Production of Excess Interleukin-3 and Granulocyte-Macrophage Colony-Stimulating Factor in Mice: A Novel Model for Chronic Myelogenous Leukemia. *Blood* **92**, 3829–3840 (1998).

754. Levine, R. L. *et al.* Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. *Cancer Cell* **7**, 387–397 (2005).

755. Baxter, E. J. *et al.* Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders. *Lancet* **365**, 1054–1061 (2005).

756. Lacout, C. *et al.* JAK2V617F expression in murine hematopoietic cells leads to MPD mimicking human PV with secondary myelofibrosis. *Blood* **108**, 1652–1660 (2006).

757. Watanabe-Okochi, N. *et al.* AML1 mutations induced MDS and MDS/AML in a mouse BMT model. *Blood* **111**, 4297–4308 (2008).

758. Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).

759. Mizuki, M. *et al.* Suppression of myeloid transcription factors and induction of STAT response genes by AML-specific Flt3 mutations. *Blood* **101**, 3164–3173 (2003).

760. Tyner, J. W. *et al.* Functional genomic landscape of acute myeloid leukaemia. *Nature* **562**, 526–531 (2018).

761. Brunetti, L., Gundry, M. C. & Goodell, M. A. New insights into the biology of acute myeloid leukemia with mutated NPM1. *International Journal of Hematology* vol. 110 150–160 (2019).

762. Ortmann, C. A. *et al.* Effect of Mutation Order on Myeloproliferative Neoplasms. *N. Engl. J. Med.* **372**, 601–612 (2015).

763. Morita, K. *et al.* Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nat. Commun.* **11**, 1–17 (2020).

764. Potter, N. *et al.* Single cell analysis of clonal architecture in acute myeloid leukaemia. *Leukemia* **33**, 1113–1123 (2019).

765. Paguirigan, A. L. *et al.* Single-cell genotyping demonstrates complex clonal diversity in acute myeloid leukemia. *Sci. Transl. Med.* **7**, 281re2 (2015).

766. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–74 (2011).

767. Valk, P. J. M. *et al.* Prognostically Useful Gene-Expression Profiles in Acute Myeloid Leukemia. *N. Engl. J. Med.* **350**, 1617–1628 (2004).

768. Wouters, B. J. *et al.* Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome. *Blood* **113**, 3088–3091 (2009).

769. Taskesen, E. *et al.* Prognostic impact, concurrent genetic mutations, and gene expression features of AML with CEBPA mutations in a cohort of 1182 cytogenetically normal AML patients: further evidence for CEBPA double mutant AML as a distinctive disease entity. *Blood* **117**, 2469–2475 (2011).

770. Verhaak, R. G. W. *et al.* Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): Association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. *Blood* **106**, 3747–3754 (2005).

771. Lin, F. T., MacDougald, O. A., Diehl, A. M. & Lane, M. D. A 30-kDa alternative translation product of the CCAAT/enhancer binding protein α message: Transcriptional activator lacking antimitotic activity. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 9606–9610 (1993).

772. Avellino, R. & Delwel, R. Expression and regulation of C/EBPα in normal myelopoiesis and in malignant transformation. *Blood* vol. 129 2083–2091 (2017).

773. Hasemann, M. S. *et al.* C/EBPα Is Required for Long-Term Self-Renewal and Lineage Priming of Hematopoietic Stem Cells and for the Maintenance of Epigenetic Configurations in Multipotent Progenitors. *PLoS Genet.* **10**, (2014).

774. Amann, J. M. *et al.* ETO, a Target of t(8;21) in Acute Leukemia, Makes Distinct Contacts with Multiple Histone Deacetylases and Binds mSin3A through Its Oligomerization Domain. *Mol. Cell. Biol.* **21**, 6470–6483 (2001).

775. Wang, J., Hoshino, T., Redner, R. L., Kajigaya, S. & Liu, J. M. ETO, fusion partner in t(8;21) acute myeloid leukemia, represses transcription by interaction with the human N-CoR/mSin3/HDAC1 complex. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 10860–10865 (1998).

776. Zhang, J., Kalkum, M., Yamamura, S., Chait, B. T. & Roeder, R. G. E protein silencing by the leukemogenic AML1-ETO fusion protein. *Science (80-.).* **305**, 1286–1289 (2004).

777. Wang, L. *et al.* The leukemogenicity of AML1-ETO is dependent on site-specific lysine acetylation. *Science (80-.).* **333**, 765–769 (2011).

778. Ptasinska, A. *et al.* Depletion of RUNX1/ETO in t(8;21) AML cells leads to genome-wide changes in chromatin structure and transcription factor binding. *Leukemia* **26**, 1829–1841 (2012).

779. Ptasinska, A. *et al.* Identification of a dynamic core transcriptional network in t(8;21) AML that regulates differentiation block and self-renewal. *Cell Rep.* **8**, 1974–1988 (2014).

780. Pabst, T. *et al.* AML1-ETO downregulates the granulocytic differentiation factor C/EBPα in t(8;21) myeloid leukemia. *Nat. Med.* **7**, 444–451 (2001).

781. Stengel, K. R., Ellis, J. D., Spielman, C. L., Bomber, M. L. & Hiebert, S. W. Definition of a small core transcriptional circuit regulated by AML1-ETO. *Mol. Cell* **81**, 530-545.e5 (2021).

782. Loke, J. *et al.* C/EBPα overrides epigenetic reprogramming by oncogenic transcription factors in acute myeloid leukemia. *Blood Adv.* **2**, 271–284 (2018).

783. Melki, J. R., Vincent, P. C. & Clark, S. J. Concurrent DNA hypermethylation of multiple genes in acute myeloid leukemia. *Cancer Res.* **59**, 3730–3740 (1999).

784. Cameron, E. E., Baylin, S. B. & Herman, J. G. p15INK4B CpG Island Methylation in Primary Acute Leukemia Is Heterogeneous and Suggests Density as a Critical Factor for Transcriptional Silencing. *Blood* **94**, 2445–2451 (1999).

785. Rampal, R. *et al.* DNA Hydroxymethylation Profiling Reveals that WT1 Mutations Result in Loss of TET2 Function in Acute Myeloid Leukemia. *Cell Rep.* **9**, 1841–1855 (2014).

786. Ko, M. *et al.* Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature* **468**, 839 (2010).

787. Yamazaki, J. *et al.* Efects of TET2 mutations on DNA methylation in chronic myelomonocytic leukemia. *Epigenetics* **7**, 201–207 (2012).

788. Rasmussen, K. D. *et al.* Loss of TET2 in hematopoietic cells leads to DNA hypermethylation of active enhancers and induction of leukemogenesis. *Genes Dev.* **29**, 910–922 (2015).

789. Glass, J. L. *et al.* Epigenetic identity in AML depends on disruption of nonpromoter regulatory elements and is affected by antagonistic effects of mutations in epigenetic modifiers. *Cancer Discov.* **7**, 868–883 (2017).

790. Yan, X. J. *et al.* Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. *Nat. Genet.* **43**, 309–317 (2011).

791. Russler-Germain, D. A. *et al.* The R882H DNMT3A Mutation Associated with AML Dominantly Inhibits Wild-Type DNMT3A by Blocking Its Ability to Form Active Tetramers. *Cancer Cell* **25**, 442–454 (2014).

792. Sandoval, J. E., Huang, Y. H., Muise, A., Goodell, M. A. & Reich, N. O. Mutations in the DNMT3A DNA methyltransferase in acute myeloid leukemia patients cause both loss and gain of function and differential regulation by protein partners. *J. Biol. Chem.* **294**, 4898–4910 (2019).

793. Spencer, D. H. *et al.* CpG Island Hypermethylation Mediated by DNMT3A Is a Consequence of AML Progression. *Cell* **168**, 801-816.e13 (2017).

794. Mizuno, S. I. *et al.* Expression of DNA methyltransferases DNMT1, 3A, and 3B in normal hematopoiesis and in acute and chronic myelogenous leukemia. *Blood* **97**, 1172–1179 (2001).

795. Abdel-Wahab, O. *et al.* ASXL1 Mutations Promote Myeloid Transformation through Loss of PRC2-Mediated Gene Repression. *Cancer Cell* **22**, 180–193 (2012).

796. Fisher, C. L. *et al.* Loss-of-function additional sex combs like 1 mutations disrupt hematopoiesis but do not cause severe myelodysplasia or leukemia. *Blood* **115**, 38–46 (2010).

797. Ernst, T. *et al.* Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. *Nat. Genet.* **42**, 722–726 (2010).

798. Score, J. *et al.* Inactivation of polycomb repressive complex 2 components in myeloproliferative and myelodysplastic/myeloproliferative neoplasms. *Blood* **119**, 1208–1213 (2012).

799. Kempf, J. M. *et al.* Loss-of-function mutations in the histone methyltransferase EZH2 promote chemotherapy resistance in AML. *Sci. Rep.* **11**, 1–13 (2021).

800. Shi, J. *et al.* The Polycomb complex PRC2 supports aberrant self-renewal in a mouse model of MLL-AF9;NrasG12D acute myeloid leukemia. *Oncogene* **32**, 930 (2013).

801. Grossmann, V. *et al.* Whole-exome sequencing identifies somatic mutations of BCOR in acute myeloid leukemia with normal karyotype. *Blood* **118**, 6153–6163 (2011).

802. Li, M. *et al.* Somatic mutations in the transcriptional corepressor gene BCORL1 in adult acute myelogenous leukemia. *Blood* **118**, 5914–5917 (2011).

803. Damm, F. *et al.* BCOR and BCORL1 mutations in myelodysplastic syndromes and related disorders. *Blood* **122**, 3169–3177 (2013).

804. Schaefer, E. J. *et al.* BCOR and BCORL1 Mutations Drive Epigenetic Reprogramming and Oncogenic Signaling by Unlinking PRC1.1 from Target Genes . *Blood Cancer Discov.* (2021) doi:10.1158/2643-3230.bcd-21-0115.

805. Schoch, C. *et al.* AML with 11q23/MLL abnormalities as defined by the WHO classification: incidence, partner chromosomes, FAB subtype, age distribution, and prognostic impact in an unselected series of 1897 cytogenetically analyzed AML cases. *Blood* **102**, 2395–2402 (2003).

806. Tkachuk, D. C., Kohler, S. & Cleary, M. L. Involvement of a homolog of Drosophila trithorax by 11q23 chromosomal translocations in acute leukemias. *Cell* **71**, 691–700 (1992).

807. Ross, M. E. *et al.* Gene expression profiling of pediatric acute myelogenous leukemia. *Blood* **104**, 3679–3687 (2004).

808. Bindels, E. M. J. *et al.* EVI1 is critical for the pathogenesis of a subset of MLL-AF9-rearranged AMLs. *Blood* **119**, 5838–5849 (2012).

809. Milne, T. A. *et al.* MLL targets SET domain methyltransferase activity to Hox gene promoters. *Mol. Cell* **10**, 1107–1117 (2002).

810. Krivtsov, A. V. *et al.* H3K79 methylation profiles define murine and human MLL-AF4 leukemias. *Cancer Cell* **14**, 355–368 (2008).

811. Okada, Y. *et al.* hDOT1L links histone methylation to leukemogenesis. *Cell* **121**, 167–178 (2005).

812. Bernt, K. M. *et al.* MLL-Rearranged Leukemia Is Dependent on Aberrant H3K79 Methylation by DOT1L. *Cancer Cell* **20**, 66–78 (2011).

813. Lin, C. *et al.* AFF4, a component of the ELL/P-TEFb elongation complex and a shared subunit of MLL chimeras, can link transcription elongation to leukemia. *Mol. Cell* **37**, 429–437 (2010).

814. Mohan, M., Lin, C., Guest, E. & Shilatifard, A. Licensed to elongate: A molecular mechanism for MLL-based leukaemogenesis. *Nature Reviews Cancer* vol. 10 721–728 (2010).

815. Thol, F. *et al.* Mutations in the cohesin complex in acute myeloid leukemia: Clinical and prognostic implications. *Blood* **123**, 914–920 (2014).

816. Thota, S. *et al.* Genetic alterations of the cohesin complex genes in myeloid malignancies. *Blood* **124**, 1790–1798 (2014).

817. Tsai, C. H. *et al.* Prognostic impacts and dynamic changes of cohesin complex gene mutations in de novo acute myeloid leukemia. *Blood Cancer J.* **7**, 1–7 (2017).

818. Solomon, D. A. *et al.* Mutational inactivation of STAG2 causes aneuploidy in human cancer. *Science (80-.).* **333**, 1039–1043 (2011).

819. Smith, J. S. *et al.* Chronic loss of STAG2 leads to altered chromatin structure contributing to de-regulated transcription in AML. *J. Transl. Med.* **18**, 1–18 (2020).

820. Viny, A. D. *et al.* Dose-dependent role of the cohesin complex in normal and malignant hematopoiesis. *J. Exp. Med.* **212**, 1819–1832 (2015).

821. van der Lelij, P. *et al.* Synthetic lethality between the cohesin subunits STAG1 and STAG2 in diverse cancer contexts. *Elife* **6**, (2017).

822. Bradner, J. E., Hnisz, D. & Young, R. A. Transcriptional Addiction in Cancer. *Cell* vol. 168 629–643 (2017).

823. Yuasa, H. *et al.* Oncogenic transcription factor Evi1 regulates hematopoietic stem cell proliferation through GATA-2 expression. *EMBO J.* **24**, 1976–1987 (2005).

824. Goyama, S. *et al.* Evi-1 Is a Critical Regulator for Hematopoietic Stem Cells and Transformed Leukemic Cells. *Cell Stem Cell* **3**, 207–220 (2008).

825. Lugthart, S. *et al.* High EVI1 levels predict adverse outcome in acute myeloid leukemia: Prevalence of EVI1 overexpression and chromosome 3q26 abnormalities underestimated. *Blood* **111**, 4329–4337 (2008).

826. Ogawa, S. *et al.* Abnormal expression of Evi-1 gene in human leukemias. *Hum. Cell* **9**, 323–32 (1996).

827. Morishita, K., Parganas, E., Matsugi, T. & Ihle, J. N. Expression of the Evi-1 zinc finger gene in 32Dc13 myeloid cells blocks granulocytic differentiation in response to granulocyte colony-stimulating factor. *Mol. Cell. Biol.* **12**, 183–189 (1992).

828. Lugthart, S. *et al.* Clinical, molecular, and prognostic significance of WHO type inv(3)(q21q26.2)/t(3;3)(q21;q26.2) and various other 3q abnormalities in acute myeloid leukemia. *J. Clin. Oncol.* **28**, 3890–3898 (2010).

829. Bitter, M., Neilly, M., Le Beau, M., Pearson, M. & Rowley, J. Rearrangements of chromosome 3 involving bands 3q21 and 3q26 are associated with normal or elevated platelet counts in acute nonlymphocytic leukemia. *Blood* **66**, 1362–1370 (1985).

830. Suzukawa, K. *et al.* Identification of a breakpoint cluster region 3' of the ribophorin I gene at 3q21 associated with the transcriptional activation of the EVI1 gene in acute myelogenous leukemias with inv(3)(q21q26). *Blood* **84**, 2681–2688 (1994).

831. Morishita, K. *et al.* Activation of EVI1 gene expression in human acute myelogenous leukemias by translocations spanning 300-400 kilobases on chromosome band 3q26. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 3937–3941 (1992).

832. Yamazaki, H. *et al.* A remote GATA2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating EVI1 expression. *Cancer Cell* **25**, 415–427 (2014).

833. Shi, J. *et al.* Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation. *Genes Dev.* **27**, 2648–2662 (2013).

834. Dhanasekaran, R. *et al.* The MYC oncogene — the grand orchestrator of cancer growth and immune evasion. *Nat. Rev. Clin. Oncol. 2021 191* **19**, 23–36 (2021).

835. Abraham, B. J. *et al.* Small genomic insertions form enhancers that misregulate oncogenes. *Nat. Commun.* **8**, 1–13 (2017).

836. Rahman, S. *et al.* Activation of the LMO2 oncogene through a somatically acquired neomorphic promoter in T-cell acute lymphoblastic leukemia. *Blood* **129**, 3221–3226 (2017).

837. Holliday, R. The inheritance of epigenetic defects. *Science (80-.).* **238**, 163–170 (1987).

838. Horsthemke, B. Epimutations in human disease. *Current Topics in Microbiology and Immunology* vol. 310 45–59 (2006).

839. Mujahed, H. *et al.* AML displays increased CTCF occupancy associated with aberrant gene expression and transcription factor binding. *Blood* **136**, 339–352 (2020).

840. Wang, A. J., Han, Y., Jia, N., Chen, P. & Minden, M. D. NPM1c impedes CTCF functions through cytoplasmic mislocalization in acute myeloid leukemia. *Leukemia* **34**, 1278–1290 (2020).

841. Oey, H. & Whitelaw, E. On the meaning of the word 'epimutation'. *Trends in Genetics* vol. 30 519–520 (2014).

842. Hitchins, M. P. *et al.* Inheritance of a Cancer-Associated MLH1 Germ-Line Epimutation. *N. Engl. J. Med.* **356**, 697–705 (2007).

843. Campos, E. I., Stafford, J. M. & Reinberg, D. Epigenetic inheritance: Histone bookmarks across generations. *Trends in Cell Biology* vol. 24 664–674 (2014).

844. Wouters, B. J. *et al.* Distinct gene expression profiles of acute myeloid/T-lymphoid leukemia with silenced CEBPA and mutations in NOTCH1. *Blood* **110**, 3706–3714 (2007).

845. Jost, E. *et al.* Epimutations mimic genomic mutations of DNMT3A in acute myeloid leukemia. *Leukemia* **28**, 1227–1234 (2014).

846. Li, S. *et al.* Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nat. Med.* **22**, 792–9 (2016).

CHAPTER 2

Induced cell-autonomous neutropenia systemically perturbs hematopoiesis in *Cebpa* enhancer-null mice

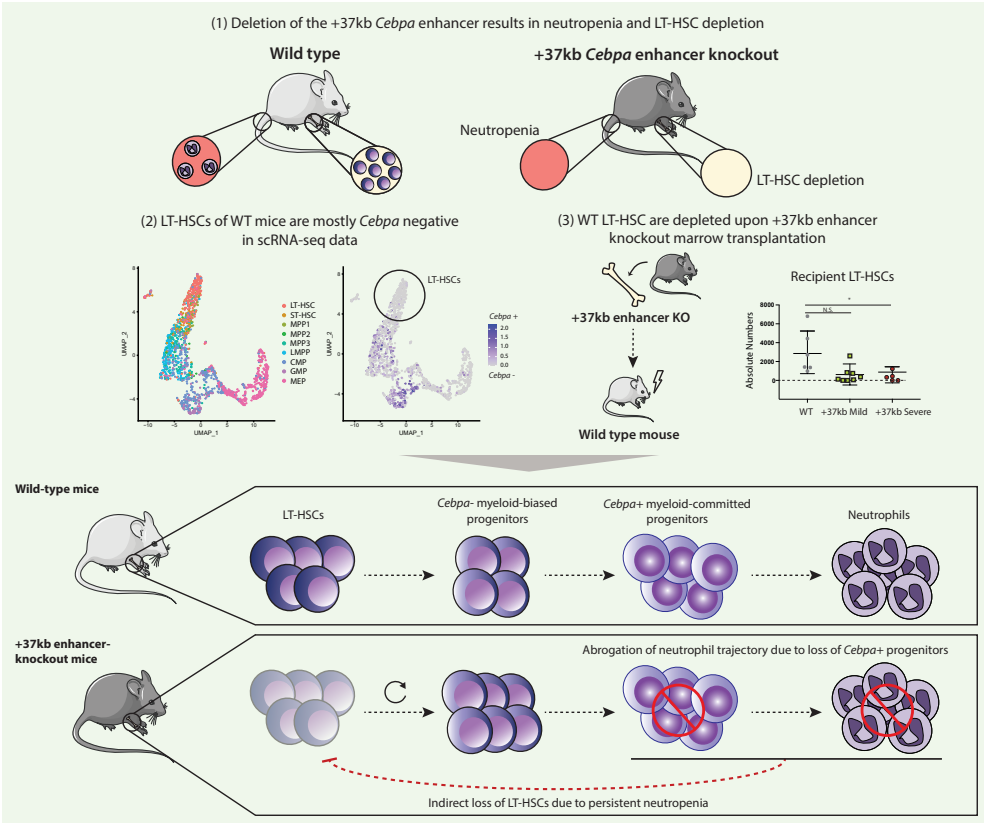
Roberto Avellino[1,2,3,*], **Roger Mulet-Lazaro[1,2,*],** Marije Havermans[1,2], Remco Hoogenboezem[1], Leonie Smeenk[1,2], Nathan Salomonis[4], Rebekka K. Schneider[1,2,5,6], Elwin Rombouts[1], Eric Bindels[1], Lee Grimes[4], Ruud Delwel[1,2]

[1] Department of Hematology, Erasmus University Medical Center, Rotterdam, The Netherlands.
[2] Oncode Institute, Erasmus University Medical Center, Rotterdam, The Netherlands.
[3] Department of Immunology, Weizmann Institute, Rehovot 76100, Israel.
[4] Division of Experimental Hematology and Cancer Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio.
[5] Current address: Department of Developmental Biology, Erasmus MC, Rotterdam, the Netherlands
[6] Institute for Biomedical Engineering, Department of Cell Biology, Rheinisch-Westfälische Technische Hochschule Aachen University, Aachen, Germany.

* These authors contributed equally to this work

Running title: Neutropenia causes LT-HSC loss and dysplasia

Blood Adv. 2022 Mar 8;6(5):1406-1419. doi: 10.1182/bloodadvances.2021005851.
PMID: 34814180


(1) Deletion of the +37kb Cebpa enhancer results in neutropenia and LT-HSC depletion
Wild type
+37kb Cebpa enhancer knockout
Neutropenia
LT-HSC depletion
(2) LT-HSCs of WT mice are mostly Cebpa negative in scRNA-seq data
LT-HSC
ST-HSC
MPP1
MPP2
MPP3
LMPP
CMP
GMP
MEP
LT-HSCs
Cebpa +
Cebpa -
UMAP_1
UMAP_2
(3) WT LT-HSC are depleted upon +37kb enhancer knockout marrow transplantation
+37kb enhancer KO
Wild type mouse
Recipient LT-HSCs
Absolute Numbers
N.S.
WT
+37kb Mild
+37kb Severe
Wild-type mice
LT-HSCs
Cebpa- myeloid-biased progenitors
Cebpa+ myeloid-committed progenitors
Neutrophils
+37kb enhancer-knockout mice
Abrogation of neutrophil trajectory due to loss of Cebpa+ progenitors
Indirect loss of LT-HSCs due to persistent neutropenia

ABSTRACT

The transcription factor C/EBPa initiates the neutrophil gene expression program in the bone marrow. Knockouts of the *Cebpa* gene or its +37kb enhancer in mice show two major findings: (1) neutropenia in bone marrow and blood; (2) decrease in long-term hematopoietic stem cell (LT-HSC) numbers. Whether the latter finding is cell autonomous (intrinsic) to the LT-HSCs or an extrinsic event exerted on the stem cell compartment remained an open question. Flow cytometric analysis of the *Cebpa* +37kb enhancer knockout model revealed that the reduction in LT-HSC numbers observed was proportional to the degree of neutropenia. Single cell transcriptomics of wild type mouse bone marrow showed that *Cebpa* is predominantly expressed in early myeloid-biased progenitors, but not in LT-HSCs. These observations suggest that the negative effect on LT-HSCs is an extrinsic event caused by neutropenia. We transplanted whole bone marrows from +37kb enhancer-deleted mice and found that 40% of the recipient mice acquired full blown neutropenia with severe dysplasia and a significant reduction in the total LT-HSC population. The other 60% showed initial signs of myeloid differentiation defects and dysplasia when they were sacrificed, suggesting they were in an early stage of the same pathological process. This phenotype was not seen in mice transplanted with wild type bone marrow cells. Altogether, these results indicate that *Cebpa*-enhancer deletion causes cell autonomous neutropenia, which reprograms and disturbs the quiescence of HSCs, leading to a systemic impairment of the hematopoietic process.

KEY POINTS

- *Cebpa* activates granulocytic differentiation in early myeloid biased progenitors but not in LT-HSCs during steady state hematopoiesis
- Unresolved neutropenia caused by *Cebpa* +37kb enhancer deletion disturbs the LT-HSC pool and leads to severe bone marrow dysplasia

INTRODUCTION

Pioneering transplantation studies showed that long-term hematopoietic stem cells (LT-HSCs) possess a multi-lineage differentiation potential towards all hematopoietic lineages after myeloablative conditioning [1,2]. Bone marrow differentiation occurs along a continuum of cellular states, progressing in a trajectory from LT-HSC towards cell-lineage specific progenitors [3–6]. Hematopoietic stem and progenitor cells (HSPCs) forming these trajectories are interconnected by transcription factor networks that drive differentiation [7]. Despite their role in differentiation, it remains unclear how transcription factors protect HSCs from exhaustion to preserve bone marrow integrity [8].

The myeloid lineage transcription factor C/EBP*α,* encoded by *Cebpa*, has been studied extensively to understand its role in myeloid differentiation. The expression of *Cebpa* in myeloid cells is specifically controlled by the +37 kb enhancer (+42 kb in humans) [9,10]. Genetic knockout of *Cebpa* [11] or of its +37kb enhancer [9,12] *in vivo* (both referred to as *Cebpa* null mice) causes neutropenia concomitant with reduced LT-HSC numbers. These studies describe C/EBP*α* as one of the major myeloid regulators that interconnects HSCs with myeloid progenitors. In addition, it has been suggested that C/EBP*α* has a dual role in LT-HSCs: maintaining LT-HSC quiescence by repressing the self-renewal [13] and proliferative [14] gene expression programs, while simultaneously priming early myeloid genes [11].

One major technical limitation in these studies is the low resolving power of the technologies used to study rare cell types such as LT-HSCs. Bulk genome-wide transcriptomics measures gene expression signatures at the population level, thereby masking the presence of any rare and transient cell state of physiological importance in the bone marrow. This limitation has been overcome by high-resolution single cell technologies combined with lineage tracing or *in vivo* barcoding. Emerging findings from studies in native hematopoiesis [4,15,16] place LT-HSCs as a separate and an occasional contributing entity to myelopoiesis in contrast with a continuous HSC to myeloid state, therefore questioning the role of *Cebpa* as a myeloid priming factor in HSCs.

Here we investigated whether LT-HSC loss in *Cebpa* null mice is the cause or consequence of neutropenia. We hypothesize that either (1) LT-HSCs harboring an active *Cebpa* locus are lost upon enhancer deletion, leading to myeloid trajectory shutdown and ultimately neutropenia or, (2) myeloid-biased progenitors expressing *Cebpa* are lost upon enhancer deletion, causing neutropenia, which systematically disturbs and depletes the LT-HSC pool. To address this question, we combined previously published single-cell datasets from wild type bone marrows with bulk-cell transcriptomics from the +37kb enhancer-deleted mice. Furthermore, we transplanted *Cebpa* enhancer-deleted cells to study the possible systemic effects on hematopoiesis of the host. Using these approaches we conclude that LT-HSCs do not express detectable levels of *Cebpa* in unperturbed hematopoiesis and *Cebpa* null induced neutropenia systemically disturbs LT-HSC quiescence, leading to HSC depletion, bone marrow hypocellularity and severe dysplasia.

MATERIALS AND METHODS

RNA sequencing

Total sample RNA was extracted using Trizol with Genelute LPA (Sigma) as a carrier and SMARTer Ultra Low RNA kit for Illumina Sequencing (Clontech) was used for cDNA synthesis according to the manufacturer's protocol. cDNA was sheared with the Covaris device and processed according to the TruSeq RNA Sample Preparation v2 Guide (Illumina). Amplified sample libraries were subjected to paired-end sequencing (2 x 75 bp) and aligned against mm10 using TopHat [17]. Gene expression levels were quantified by the fragments per kilobase of exon per million fragments mapped (FPKM) statistic as calculated by Cufflinks [18] in the RefSeq Transcriptome database [19]. Read counts were determined with HTSeq-count [20] and subsequently used for differential expression analysis in DESeq2 [21], with default parameters, in the R environment. Multiple testing correction was performed by the Benjamini-Hochberg procedure on the calculated p-values to control the False Discovery Rate (FDR).

For gene set enrichment analysis (GSEA), a ranking metric was defined for each gene as the log10 of the adjusted p-value calculated by DESeq2 with the sign of the log2 fold change. The ranked gene list was tested against a customized version of the C2 MSigDB collection, incorporating datasets on HSC quiescence from the literature [3,22–24] (Supplementary Table 1).

Single-cell RNA sequencing

A compendium of previously published mouse bone marrow single-cell RNA-Seq datasets (Fluidigm C1 platform) was assembled to evaluate the expression of key progenitor genes of interest described [25–27] (available at: http://www.altanalyze.org/ICGS/Public/Mm-Grimes-Fluidigm-Panorama/User.php). Specifically, wild type in vivo mouse SLAM, LSK, GMP, CMP and lineage-negative Sca1+ CD117+ cells from bone marrow were selected, with labels derived from the noted prior studies (GSE70245, GSE141472). Data were analyzed with RSEM to estimate TPM for all genes as previously described [28]. The gene expression data of selected genes were visualized in the python package *plotly* or GraphPad Prism, represented as log2 TPM values.

For the analysis in Figure S1, published single-cell RNA-seq data of HSPCs from the bone marrow of 10 female 12-week-old C57BL/6 mice were retrieved [29]. Raw counts were downloaded from GEO database (GSE81682), with labels derived from the broad gating strategy used in the original study (available here: http://blood.stemcells.cam.ac.uk/data/all_cell_types.txt). Data were imported and processed using the Seurat R package [30]. Cells with fewer than 200000 total counts, fewer than 4000 detected genes or more than 10% mitochondrial reads were excluded. Expression data were log-normalized and the 5000 most variable features were selected for dimensionality reduction with UMAP. *Cebpa* expression was projected on the UMAP and compared to the expression of other selected genes (*Mecom*, *Hlf)* to investigate its association with different cell populations.

Mice and transplantation procedures

The strains of +37kb enhancer-1.2kb and +37kb enhancer-1.15kb deleted mice were generated using zygotes derived from C57/BL6 mice by CRISPR/Cas9 editing and maintained as previously described [9]. The CRISPR/Cas9 single guide RNAs were directed against the +37kb enhancer of *Cebpa*, using the following sequences:

5' : TGAAGCCTACACTACTTTGT and AGAGGTAGGAACTCCATTCC

3' : AGAGCCTCGCTCAAGCCCAT and TTGAGACATCTGGTAACCTT

Recipient mice were exposed to a 5.5 Gy of gamma radiation. Given that the native +37kb^{HOM} mouse exhibits 3-5 fold increase in the cKIT+ Lineage negative progenitor fraction of the bone marrow, we transplanted one million of WT CD34.2 total bone marrow cells and 250,000 of HOM CD34.2 cells to compensate for the fold difference. Non-transplanted mice were sacrificed for FACS analysis between 4-8 weeks of age. For transplantation experiments, bone marrow from 4 weeks old wild type or +37kb enhancer-deleted CD45.2 mice were harvested in PBS/5%FCS and injected intravenously in tails of (CD45.1) female mice. All mice were sacrificed in a CO_2 chamber. Animal studies were approved by the Animal Welfare/ Ethics Committee of the EDC in accordance with legislation in the Netherlands (approval No. EMC 2067, 2714, 2892, 3062).

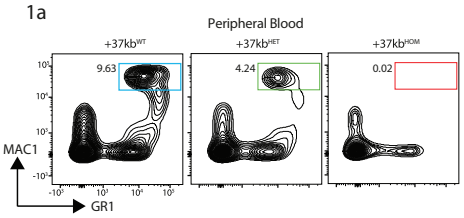
Flow-cytometry and sorting

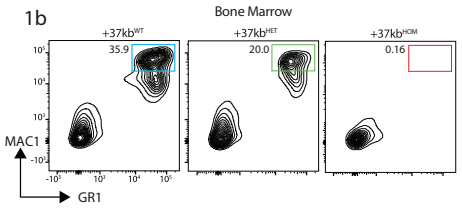
Flow cytometry was carried out on the LSRII and the FACSCanto II (BD Biosciences). FACS AriaIII (BD Biosciences) was used for cell sorting, using the following fluorescent antibodies: Markers for mature hematopoietic cells [CD11B APC/GR1-FITC/B220-PE/CD3 PB]; Pan-hematopoietic marker [CD45.1 PE; CD45.2 APC-CY7]; LIN biotinylated cocktail [CD11B, GR-1, B220 and CD3] streptavidin-pacific orange; LSK [cKIT-APC/SCA1-PB]; LT-HSCs [CD48-FITC/ CD150-PE-CY7]. Lineage negative selection was carried out using a cocktail of antibodies targeting antigens expressed on mature hematopoietic cells including CD11B, GR-1, B220 and CD3. All antibodies were purchased from BD Biosciences or Biolegend (Supplementary Table 2). The LSK population was gated from Live (DAPI$^-$), CD45+, Lineage$^-$cKIT$^+$ SCA-1$^+$ bone marrow cells and the sorted LSK fraction was collected in 500µl PBS with 5% FCS, spun down and re-suspended in 800µl of Trizol and used for RNA-seq.

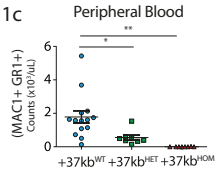
RESULTS

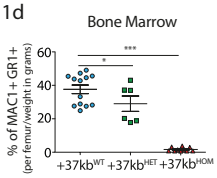
Neutropenia results in LT-HSC number reduction in +37kb Cebpa-enhancer-deleted mice

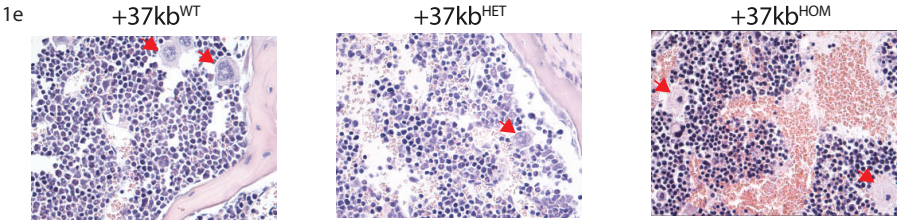
To study the causal relationship of neutropenia and loss of LT-HSCs, we investigated the allelic dosage effect of *Cebpa* enhancer deletion on the numbers of neutrophils and LT-HSCs using *Cebpa* +37kb enhancer heterozygous (+37kb^{HET}) deleted and +37kb homozygous (+37kb^{HOM}) deleted mouse strains [9]. The neutrophil (Mac1+Gr1+) frequency and absolute numbers in the peripheral blood and bone marrow of the +37kb^{HET} and the +37kb^{HOM} mice correlated with their mono-allelic and bi-allelic enhancer deletion, respectively (Figure 1a-1d). Therefore, 50% enhancer activity reduced the neutrophil output approximately by half (median frequency: 44.6%) (Figure 1c), while a full enhancer deletion completely abrogated neutrophil production (median frequency 0%). The reduction of neutrophils occurring in the +37kb^{HOM} mice was confirmed by hematoxylin and eosin (H&E) staining of bone marrow sections (Figure 1e) and by the decreased frequency of S100A8 cells (Figure 1f). To investigate whether enhancer deletion affects the LT-HSC pool, we studied the LT-HSC population using the SLAM CD150 and CD48 markers (LT-HSCs: LSK/CD150+CD48-) in the +37kb^{HET} and +37kb^{HOM} mice. The LT-HSC pool was not affected in the +37kb^{HET} mice despite the 50% enhancer dosage (Figure 1g and 1h). Although there was a near 2-fold reduction in LT-HSC numbers, this difference was not statistically significant. In contrast, a 10-fold reduction of LT-HSC numbers at 100% enhancer dosage reduction was observed in the +37kb^{HOM} mice (Figure 1g and 1h). The fact that mono-allelic enhancer deletion does not affect LT-HSC numbers, but reduces the neutrophil compartment by half, suggests that *Cebpa* only becomes expressed in neutrophil-primed progenitors and is inactive in LT-HSCs. Thus, the quantitative changes in the LT-HSC pool of the +37kb^{HOM} mice could be explained by an indirect effect of complete ablation of neutrophils, whereas partial neutropenia in the +37kb^{HET} is insufficient to inflict these changes. Alternatively, myeloid progenitors may be more sensitive to reduced *C/EBPa* levels than LT-HSCs, which would only become depleted upon biallelic deletion of the +37 kb enhancer.

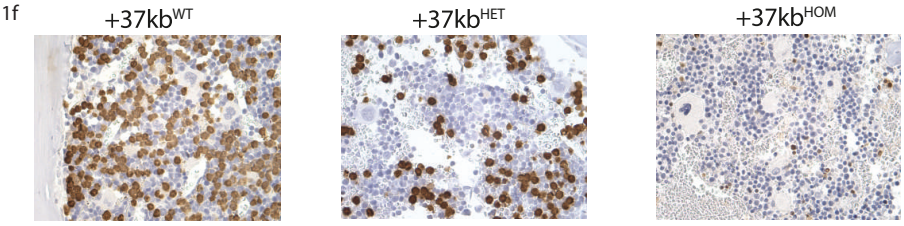

1a
Peripheral Blood
+37kb^WT
+37kb^HET
+37kb^HOM
9.63
4.24
0.02
MAC1
GR1

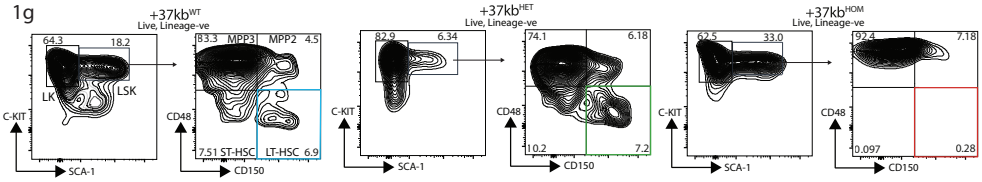

1b
Bone Marrow
+37kb^WT
+37kb^HET
+37kb^HOM
35.9
20.0
0.16
MAC1
GR1

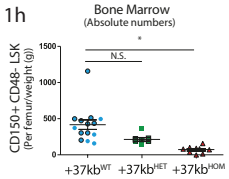

1c
Peripheral Blood
(MAC1+ GR1+)
Counts (x10^3/uL)
+37kb^WT +37kb^HET +37kb^HOM


1d
Bone Marrow
% of MAC1+ GR1+
(per femur/weight in grams)
+37kb^WT +37kb^HET +37kb^HOM


1e
+37kb^WT
+37kb^HET
+37kb^HOM


1f
+37kb^WT
+37kb^HET
+37kb^HOM


1g
+37kb^WT
Live, Lineage-ve
64.3
18.2
LK
LSK
C-KIT
SCA-1
83.3
MPP3
MPP2
4.5
CD48
7.51 ST-HSC
LT-HSC 6.9
CD150
+37kb^HET
Live, Lineage-ve
82.9
6.34
74.1
6.18
10.2
7.2
+37kb^HOM
Live, Lineage-ve
62.5
33.0
92.4
7.18
0.097
0.28


1h
Bone Marrow
(Absolute numbers)
CD150+ CD48- LSK
(Per femur/weight (g))
N.S.
+37kb^WT +37kb^HET +37kb^HOM

Cebpa is predominantly expressed in early myeloid-biased progenitors

To study *Cebpa* expression along differentiation from LT-HSCs towards multipotent and myeloid progenitors, we utilized single-cell RNA sequencing datasets generated by a Fluidigm-based platform. In total, 1,110 bone marrow cells were analyzed (27 libraries, ~50 cells per library and >2 million reads/cell on average) and assigned to a specific bone marrow cell type. The identified cellular states were classified into LT-HSC (SLAM) population, HSPC-1 and HSPC-2 (both LSK), and multi-lineage progenitors, which constitute a mixture of bi-potent and uni-potent restricted myeloid progenitors, as previously described [25–27]. To investigate at what differentiation stage *Cebpa* becomes expressed, we first used known gene markers associated with HSC quiescence, namely *Hlf* [3] and *Mecom* [31], and early myeloid differentiation (*Elane*, *Mpo*) (Figure 2a). We found a positive correlation between *Hlf* and *Mecom*, while *Cebpa* negatively correlated with *Hlf*/*Mecom* (Figure 2b). Accordingly, *Cebpa* expression was almost negligible in LT-HSCs (detected in 0.8 % of LT-HSCs), while it appeared in a subset of HSPCs expressing early myeloid lineage genes, indicating priming of the myeloid lineage at very early stages of differentiation (Figure 2c). The detection limit of this technique is 0.25 transcripts per million (TPM), which could possibly exclude cells with very low but present levels of *Cebpa*. However, these results were confirmed in another dataset [29] generated by a different single cell sequencing strategy (Figure S1). Altogether, our single cell analysis does not support a cell-autonomous role for *Cebpa* in LT-HSCs. Thus, the reduction in LT-HSC numbers observed in the *Cebpa* null mice seems to occur systemically as a consequence of neutropenia.

Figure 1. *Cebpa* enhancer deletion causes neutropenia and reduction in LT-HSCs of +37kb^{HOM} bone marrows. (a-b) Representative flow cytometry plots showing Mac1+Gr1+ myeloid cell populations in peripheral blood (a) and bone marrow (b) of +37kb^{WT} (blue), +37kb^{HET} (green) and +37kb^{HOM} (red) mice; (c) Absolute numbers of Mac1+Gr1+ cells calculated from total peripheral blood counts of 37kb^{WT} (blue), +37kb^{HET} (green) and +37kb^{HOM} (red) mice; (d) Relative numbers of total Mac1+Gr1+ cells calculated from total bone marrow cell count from 1 femur, corrected for body weight in grams of each mouse. (e) Hematoxylin and Eosin (H&E) staining of representative cross sections showing bone marrow architecture (left), identifying megakaryocytes [right (arrows)] in +37kb^{WT} and +37kb^{HOM} mice. (f) S100A8 Immunohistochemical staining of representative bone marrow cross sections from 37kb^{WT}and +37kb^{HOM} mice. (g) Representative flow cytometry plots of LK, LSK, LT-HSC, ST-HSC, MPP3, MPP2 cell populations in bone marrow of 37kb^{WT}, +37kb^{HET} and +37kb^{HOM} mice. (h) Absolute numbers of CD150+CD48- LT-HSCs calculated from total bone marrow cell counts from 1 femur. All data are represented as mean +/- SD. Statistical significance was calculated using a Student's t-test: N.S. = not significant; pvalue <0.05 (*); pvalue <0.005 (**); pvalue <0.0005 (***); SD = standard deviation. LK: Lineage-, cKit+; LSK: Lineage- Sca1+ cKit+; LT-HSC: long-term hematopoietic stem cell; ST-HSC: short-term hematopoietic stem cell; MPP: multipotent progenitor; +37kb^{WT}: wild type +37kb enhancer; +37kb^{HET}: heterozygous +37kb enhancer deletion; +37kb^{HOM} homozygous +37kb enhancer deletion.

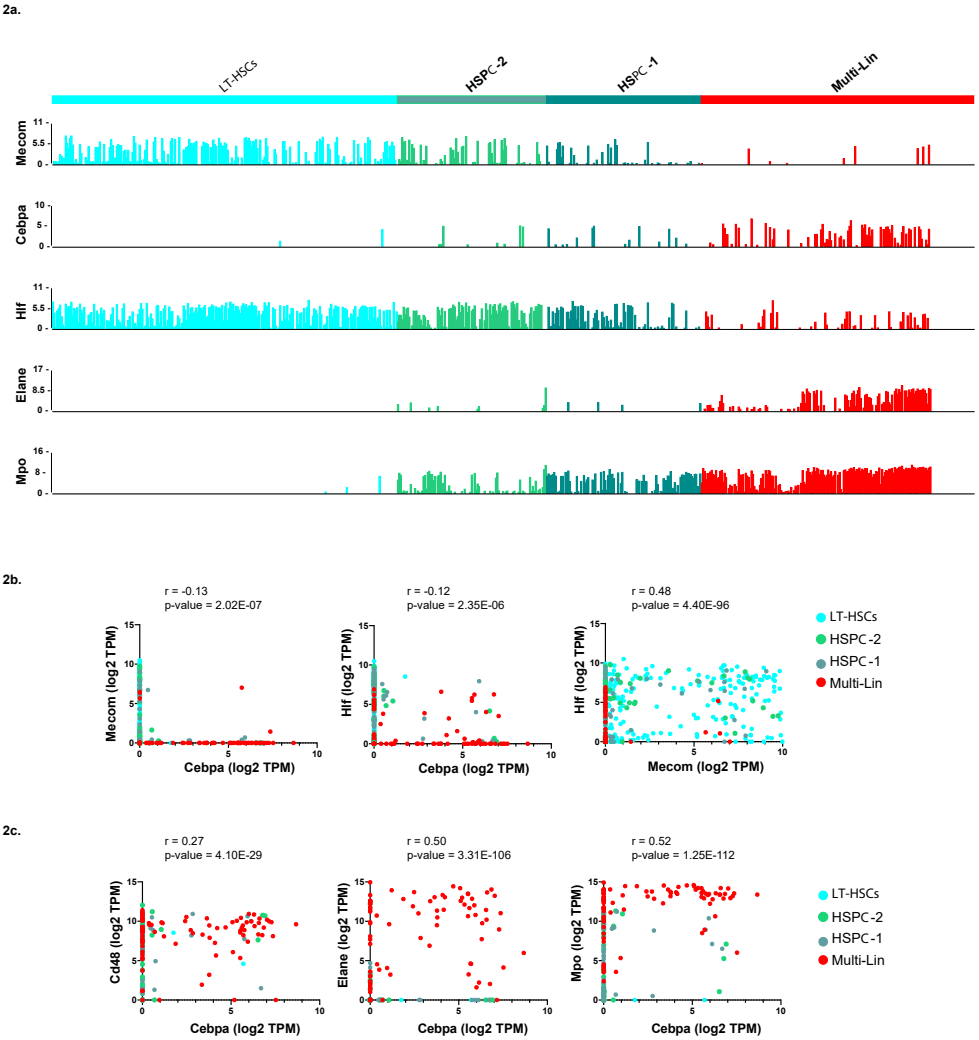


Figure 2. Single cell RNA sequencing data in wild type bone marrows exclude the presence of *Cebpa*-expressing LT-HSCs. (a) Bar plot of single cell expression of *Cebpa* in different progenitor populations, alongside genes involved in LT-HSC quiescence (*Mecom, Hlf*) or myeloid differentiation (*Elane, Mpo*) (b) Scatterplots of single cell gene expression showing that *Mecom* and *Cebpa* are mutually exclusive, whereas *Hlf* and *Cebpa* are occasionally co-expressed and *Mecom* frequently co-occurs with *Hlf* (c) Scatterplots of single cell gene expression showing co-expression of *Cebpa* with *Cd48* (a marker not found in LT-HSCs) and the myeloid differentiation markers *Elane* and *Mpo*. All gene expression data are presented as transcripts per million (TPM). Cells are color coded by the population they belong to. Pearson correlation coefficients and the related p-values for the pairwise gene combinations are depicted in the scatter plots.

Transcriptional programs of LT-HSC quiescence and neutrophil lineage-priming are deregulated in +37kb^{HOM} HSPCs

The results so far indicate that the enhancer deletion decreased the pool of LT-HSCs through a systemic effect in the presence of neutropenia. To study mechanisms linking the two observed events, we applied RNA sequencing on sorted HSPCs using LSK markers on +37kb^{WT} (n=3) and +37kb^{HOM} (n=3) bone marrow cells. We confirmed that +37kb enhancer deletion in HSPCs reduces *Cebpa* expression relative to wild type controls (Figure 3a). Differential expression analysis identified dysregulated genes related to neutrophil differentiation (*S100a9*, *Camp*) and HSC quiescence (*Hlf*) (Figure 3b, Supplementary Table 3). Differentially expressed genes in +37kb^{HOM} LSKs were further investigated by gene-set enrichment analysis (GSEA) (Supplementary Tables 4-5). Genes associated with neutrophil ontogeny were downregulated in +37kb^{HOM} LSK cells (Figure 3c), indicating an early block in myeloid differentiation. Interestingly, we found an early myeloid gene set (*Cpa3, Mpo, Cd48*) (Figure 3d) to be upregulated in +37kb^{HOM} HSPCs, suggesting that an early myeloid-biased population upstream of *Cebpa*-HSPCs is primed for myelopoiesis. Based on these findings we conclude that *Cebpa*-HSPCs are intermediate progenitors that link early myeloid biased HSPCs to myeloid committed progenitors and they represent the cell of origin for the induced neutropenia in *Cebpa* null mice.

To investigate pathways related to LT-HSC depletion, we retrieved datasets from published hematopoietic studies and pooled them with the MSigDB datasets [3,23,32–34]. Of the most significantly (FDR <0.025) enriched pathways, 6 gene sets were enriched for loss of HSC quiescence and exhaustion (Figure 3e). The identified gene sets (Figure 3f; Figure S2a-f) included transcription factors related to HSC dormancy (*Hlf, Mecom, Tcf15)* [3,31,35], HSC retention factors (*Ptpn21, Cxcr4*) [36,37], cluster of differentiation (Cd) markers for HSC regeneration and engraftment (*Cd81, Cd274*) [38,39], and the Polycomb chromobox proteins for HSC self-renewal regulation (*Pbx6*, *Pbx7*) (gene lists in Supplementary Table 1). Deregulation of ribosomal genes (Figure 3g) in our dataset is in accordance with loss of HSC quiescence. Thus, LT-HSC loss in the presence of neutropenia may result from quiescence exit and subsequent exhaustion. Altogether, the transcriptomic analysis of sorted LSK cells from +37kb^{HOM} mice suggests that myeloid priming in HSPCs occurs before *Cebpa* activation and that a neutrophil differentiation block in these progenitors potentially leads to HSC quiescence exit.

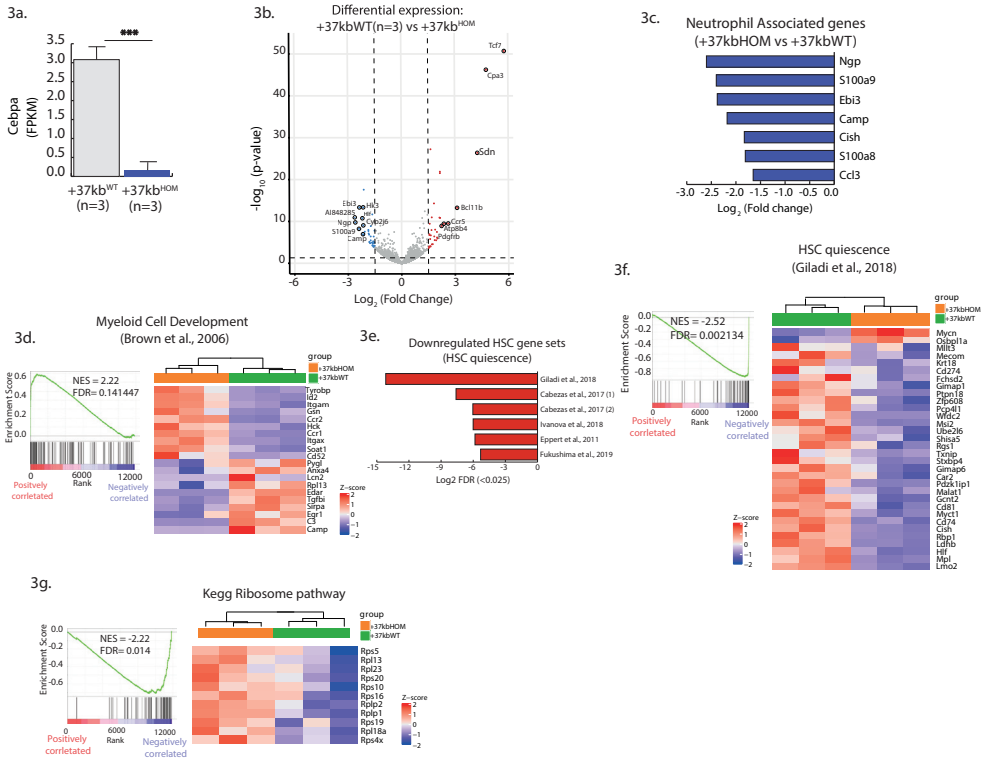


Figure 3. Transcriptome analysis reveals loss of HSPC quiescence in +37kb^{HOM} mice. (a) Volcano plot showing genes differentially expressed in HSPCs of +37kb^{HOM} (n = 3) compared to +37kb^{WT} mice (n = 3). Differentially expressed genes are represented as log2 fold change (X-axis) and log10 p-value (Y-axis). Significantly upregulated genes (log2 fold change > 1.5, p-value < 0.05) are shown in red, downregulated genes (log2 fold change < -1.5, p-value < 0.05) are shown in blue, and genes without significant differences are shown in grey. (b) Mean and SD of *Cebpa* expression in +37kb^{WT} and +37kb^{HOM} LSKs, expressed as FPKM values. (c) Bar plot showing the top downregulated neutrophil-associated genes in +37kb^{HOM} HSPCs compared to +37kb^{WT} HSPCs, presented as log2 fold change. (d) GSEA enrichment plot (left) showing upregulation (NES= 2.22; FDR <0.05) of the early myeloid-biased gene expression program in +37kb^{HOM} HSPCs. Heatmap (right) showing significant differentially expressed genes of this pathway in +37kb^{HOM} vs +37kb^{WT} HSPCs. (e) Bar plot showing downregulation of LT-HSC quiescence-related pathways in +37kb^{HOM} vs +37kb^{WT} HSPCs, expressed as -log2 transformation of FDR. (f) GSEA enrichment plot (left) showing downregulation (NES = -2.52; FDR < 0.05) of the HSC quiescence pathway from [3] in +37kb^{HOM} vs +37kb^{WT} HSPCs. Heatmap (right) of the significant differentially expressed genes in this dataset. (g) GSEA enrichment plot (left) showing downregulation (NES = -2.22; FDR <0.05) of the ribosome pathway and heatmap (right) of differentially expressed ribosomal *Rsp* and *Rlp* genes. (h) Mean and SD of *Mycn* expression in +37kb^{WT} and +37kb^{HOM} HSPCs, expressed as FPKM values. HSPC: hematopoietic stem and progenitor cells; SD: standard deviation; FDR: False discovery rate; FPKM: Fragments Per Kilobase of transcript per Million. The experiment was done in triplicates for each condition, +37kb^{WT} (n=3) and +37kb^{HOM} (n=3) and the heatmap values were calculated using Z-scores.

Transplantation of +37kb^{HOM} bone marrow into wild type recipients confirms that neutropenia is induced by the +37kb enhancer deletion

Next, we studied the systemic effects of the +37kb^{HOM} bone marrow on normal hematopoiesis in recipient mice. A sub-lethal irradiation approach was used to overcome the low survival rate caused by the weak chimerism known to occur when transplanting *Cebpa* null bone marrow cells into lethally irradiated recipients [11]. We used donor bone marrow cells from two different +37kb^{HOM} mouse strains, to ensure that the observed phenotype was not caused by an off-target effect. The strains differed in the genomic size of enhancer deletion generated by CRISPR, i.e. +37kb^{HOM} deletion of 1.15kb (CD45.2 +37kb$^{HOM\text{-}1.15kb}$ deletion) or +37kb^{HOM} deletion of 1.2kb (CD45.2 +37kb$^{HOM\text{-}1.2kb}$) (Figure S3a-S3b). Sub-lethally irradiated recipient (CD45.1) mice were transplanted with total bone marrow of CD45.2 +37kb^{WT} (N=8), CD45.2 +37kb$^{HOM\text{-}1.15kb}$ (N=7) or CD45.2 +37kb$^{HOM\text{-}1.2kb}$ (N=8) mice (Figure 4a). As expected, peripheral blood samples withdrawn twelve weeks after transplantation showed a weaker chimerism in mice transplanted with +37kb^{HOM} bone marrow compared to those transplanted with wild type bone marrow (Figure 4b). For cellular reconstitution, we determined the frequency of myeloid cells (Mac1+Gr1+), B-cells (B220+) and T-cells (CD3+) by flow cytometry for mice transplanted with +37kb^{WT} or +37kb^{HOM} bone marrow cells (Figure 4c-f, Figure S4a-c).

Populations from the T-cell, myeloid and B-cell lineages were present in the recipients from the three cohorts, indicating that hematopoiesis is functional in the recipient (Figure 4d upper panel and 4f upper panel). However, +37kb^{HOM} donor cells (Figure 4e lower panel and Figure 4f lower panel) lacked myeloid reconstitution as compared to +37kb^{WT} (Figure 4c lower panel and Figure 4d lower panel). Therefore, transplanted +37kb^{HOM} bone marrow cells in a wild type niche did not recover neutrophil differentiation, confirming that neutropenia in +37kb^{HOM} enhancer-deleted mice is cell autonomous. In the +37kb^{WT} donor cell compartment, there was a large contribution of B-cells, as previously observed in other transplantation experiments [40–42]. Of note, the presence of B-cells and T-cells from the +37kb^{HOM} donors in both the native and the transplanted mice also argues against a strict need for *Cebpa* in LT-HSCs, since they were able to survive and differentiate into the lymphoid lineage.

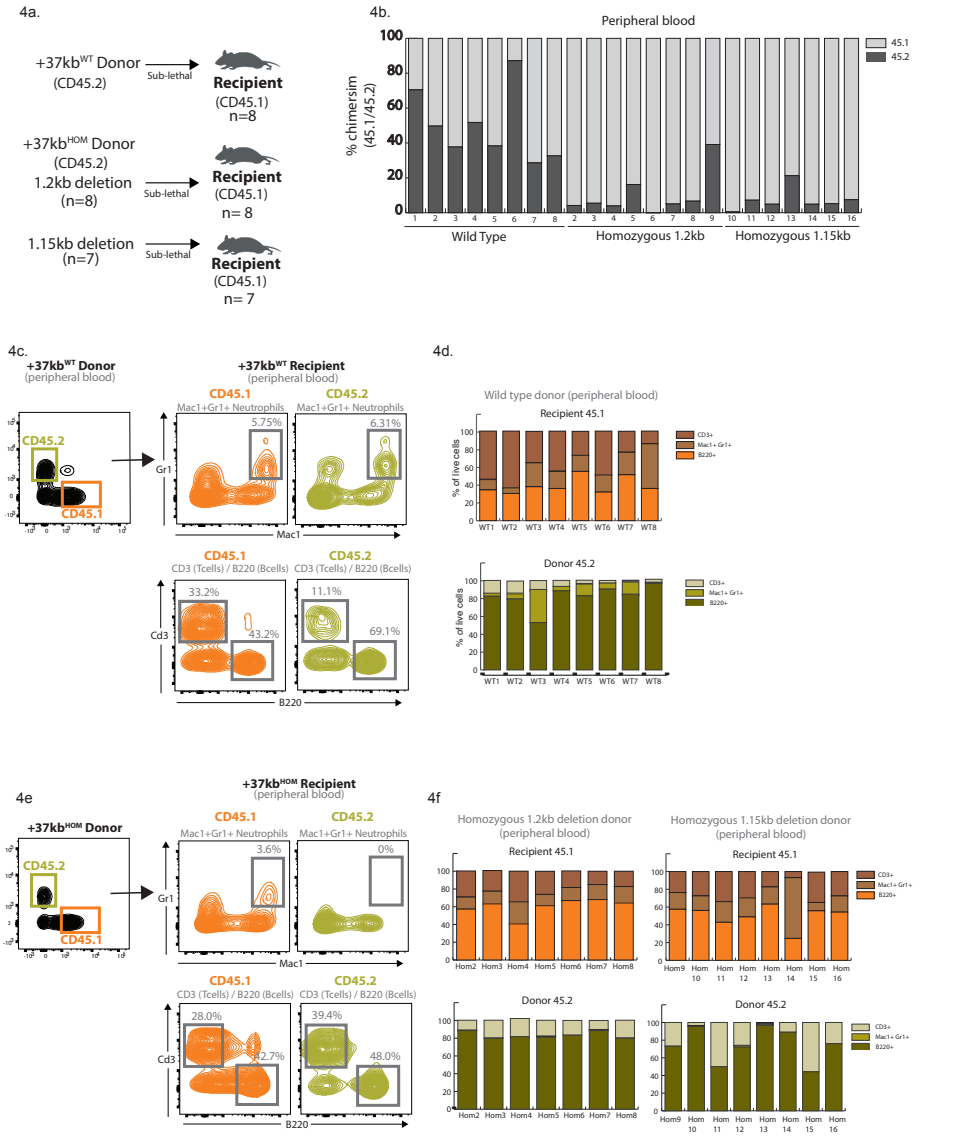


Figure 4. **Transplanted +37kb enhancer-deleted bone marrow cells show low chimerism and are neutropenic.** (a) Transplantation scheme representing donor (CD45.2) wild type, donor (CD45.2) with homozygous 1.2kb deletion and donor (CD45.2) with homozygous 1.15kb deletion in (CD45.1) recipients. Eight recipient mice per condition (n=24) were used in the experiment. One mouse transplanted with 1.2kb enhancer-deleted bone marrow cells died before it could be analyzed. (b) Bar chart showing percentage of CD45.1 and CD45.2 cell chimerism in peripheral blood twelve weeks after transplantation of wild type and homozygous bone marrow. (c) Flow cytometry contour plot of bone marrow after transplantation with +37kb^{WT} cells: Mac1+Gr1+ myeloid cells, B220+ B-cells and CD3+ T-cell populations were gated from CD45.1 and CD45.2 fractions. (d) Proportion of +37kb^{WT} Mac1+Gr1+ myeloid cells, B220+ B-cells and CD3+ T-cell populations from recipient CD45.1 (up) and donor CD45.2 (lower). (e) Flow cytometry contour plots of bone marrow after transplantation with +37kb^{HOM} cells: Mac1+Gr1+ myeloid cells, B220+ B-cells and CD3+ T-cell populations were gated from CD45.1 and CD45.2 fractions. (f) Proportion of +37kb^{HOM} Mac1+Gr1+ myeloid cells B220+ B-cells and CD3+ T-cell populations from recipient CD45.1 (upper) and donor CD45.2 (lower). Experiments were conducted on peripheral blood samples of +37kb^{WT} (n=8) and +37kb^{HOM} (n=15), drawn 12 weeks post-transplantation.

Transplanted +37kb^{HOM} bone marrow causes dysplasia and hypocellularity in recipient mice

Despite the low chimerism, the CD45.2 +37kb^{HOM} bone marrow cells persisted and survived for at least 10 months in the CD45.1 recipient, but eventually declined to almost 0% (median = 0.89%) on the day the mice were sacrificed (Figure S5a). Therefore, cells from the blood and bone marrow samples analysed were mostly derived from the recipient mice. We grouped the recipient mice based on their phenotype severity, i.e. +37kb$^{HOM\text{-}mild}$ (n=8) or +37kb$^{HOM\text{-}severe}$ (n=6, namely mice # 2,5,7,9,12 and 17). The +37kb$^{HOM\text{-}severe}$ mice represented 40% of all the transplanted mice that showed physical weakness (squinting eyes, hunched posture and social isolation), which had to be sacrificed (Figure S5b). One of the mice died before it could be analyzed. Bone marrow hypocellularity and pancytopenia were the primary hallmarks of the +37kb$^{HOM\text{-}severe}$ group (Figures 5a and S5c). Hypocellular bone marrows showed remodeling of blood vessels surrounded by a high degree of immature hematopoietic cells characterized by incomplete differentiation (Figure 5b). The +37kb$^{HOM\text{-}mild}$ group showed lower white blood cell (Figure S6a) and platelet counts (Figure S6c) compared to wild type controls, whereas hemoglobin levels were comparable to those of the +37kb^{WT} group (Figure S6b). Bone marrow cellularity in the +37kb$^{HOM\text{-}mild}$ mice varied from normocellular to hypocellular. Both the +37kb$^{HOM\text{-}severe}$ and the +37kb$^{HOM\text{-}mild}$ cohorts showed abnormal megakaryocytes with dysplastic features (Figure 5b). These data are in line with the dysplastic megakaryocytes found in the native +37kb-enhancer-deleted mice (Figure 1e, Figure S7), which might possibly explain the abnormal peripheral blood platelet counts in the recipient mice (Figure S6c). In conclusion, a significant number of mice transplanted with +37kb^{HOM} bone marrow exhibit perturbed hematopoiesis featuring hypocellularity and peripheral cytopenia, incomplete differentiation and severe dysplasia.

Acquired neutropenia leading to LT-HSC loss is recapitulated in mice transplanted with +37kb^{HOM} bone marrow

Next, we sought to study whether the differences between the +37kb$^{HOM\text{-}severe}$ and the +37kb$^{HOM\text{-}mild}$ phenotypes involve maturation defects of myeloid cells or LT-HSCs in the bone marrow of recipient mice. The +37kb$^{HOM\text{-}severe}$ transplanted mice showed a marked reduction in the numbers of recipient (CD45.1) Mac1+Gr1+ cells (Figure 5c), which was confirmed by the decrease in S100A8 protein-expressing cells (Figure 5d). In contrast, the +37kb$^{HOM\text{-}mild}$ mice showed normal neutrophil numbers compared to the wild type controls (Figure 5c). The +37kb$^{HOM\text{-}severe}$ mice also showed a reduction in the CD45.1 HSPC and LT-HSC numbers (Figures 5e and 5f), whereas the LT-HSC numbers in the +37kb$^{HOM\text{-}mild}$ mice were comparable to those of the +37kb^{WT} group. Thus, the findings in the +37kb$^{HOM\text{-}severe}$ support the notion that *Cebpa* null-induced neutropenia triggers LT-HSC loss in a cell non-autonomous manner, as observed in 37kb^{HOM} native mice.

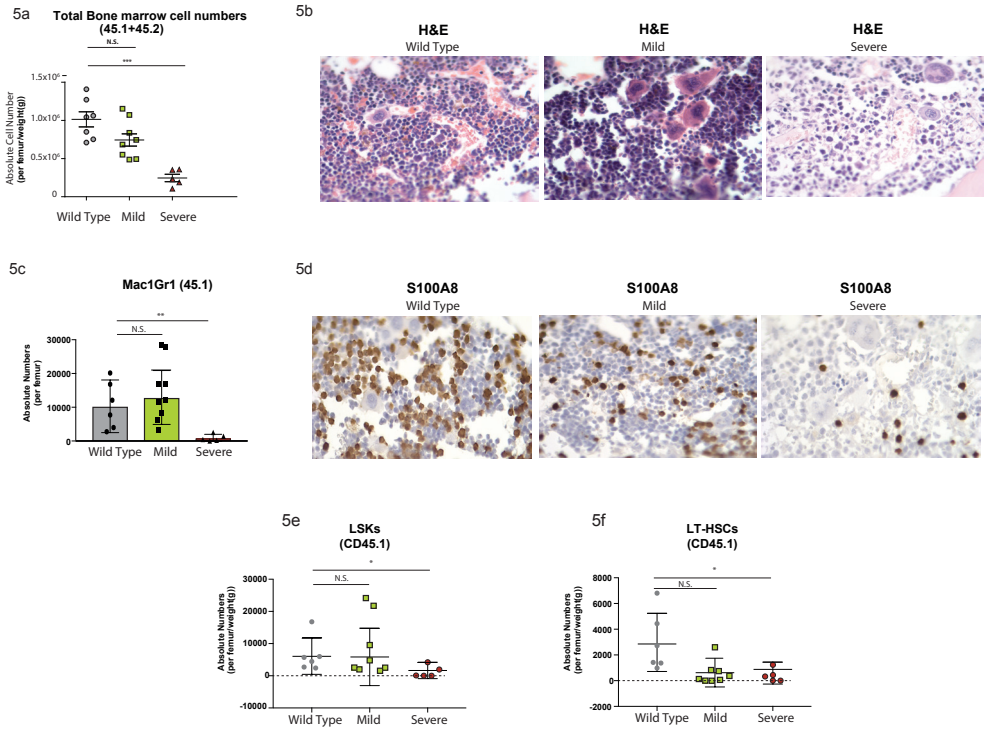


Figure 5. Systemic bone marrow perturbations impair myeloid differentiation and deplete the LT-HSC pool of recipient mice. (a) Total bone marrow cellularity per femur, corrected for body weight in grams of each mouse. The mice transplanted with homozygous bone marrow cells were divided into two subgroups based on disease severity: homozygous mild and homozygous severe. Mice transplanted with wild type bone marrows were used as controls. (b) Histological examination using hematoxylin and eosin (H&E) on processed paraffin bone marrow sections of recipient CD45.1 transplanted mice. (c) Absolute numbers of Mac1+Gr1+ neutrophils from bone marrow of CD45.1 recipient mice. (d) Immunohistochemistry of S100A8 protein expression on histological bone marrow sections of recipient CD45.1 transplanted mice (e) HSPC absolute numbers in bone marrows of CD45.1 recipient mice, calculated from LSK/lineage negative/live cells (7AAD) and corrected for total cellularity per femur and body weight (in grams) of each mouse. (f) LT-HSC absolute numbers in bone marrow of CD45.1 recipient mice, calculated from CD48-Cd150+/LSK/lineage negative/live cells (7AAD) and corrected for total cellularity per femur and body weight in grams of each mouse. All data are represented as mean +/- SD. Statistical significance was calculated using a Student's t-test: N.S. = not significant; pvalue <0.05 (*); pvalue <0.005 (**); pvalue <0.0005 (***); SD = standard deviation.

DISCUSSION

C/EBPα is indispensable for myeloid lineage formation and differentiation [13,43]. Induced genetic defects in the *Cebpa* locus of mouse models [9,12,44–46] or *CEBPA* coding mutations detected in human acute myeloid leukemia specimens (AML)[47,48], are all associated with myeloid differentiation abnormalities. The expression of *CEBPA* is also frequently repressed in leukemia by transcriptional or post-transcriptional mechanisms [8]. Notably, *CEBPA* is silenced by oncoproteins targeting the +42 kb enhancer (homologous to +37 kb in mouse), including RUNX1-RUNX1T1 [49] and *EVI1* [50]. Earlier mouse studies revealed an indispensable role for *Cebpa* in differentiating common myeloid progenitors into granulocyte-monocyte progenitors. The profound neutropenia observed in our +37 kb enhancer-deleted model confirms the absolute requirement for *Cebpa* in myelopoiesis. We observed a 3-fold increase in *Cebpb* expression in cKit+ Sca1- Cd34+ myeloid progenitors, but this was insufficient to compensate for the loss of *Cebpa.* Although C/EBPbeta can substitute for C/EBPalpha during hematopoiesis when knocked into the *Cebpa* gene locus [51], it cannot fully replace C/EBPalpha when expressed from its native locus.

Several studies hinted towards an intrinsic role for *Cebpa* in myeloid lineage priming of LT-HSCs [11,12]. However, the bulk-sequencing methods used in these studies do not meet the resolution required to dissect the heterogeneity of the HSPC compartment. Using single cell RNA sequencing, we showed that *Cebpa* is barely detectable in LT-HSCs during steady-state hematopoiesis. In contrast to previous studies, we found that *Cebpa* licenses myeloid-primed HSPCs downstream of LT-HSCs for neutrophil lineage differentiation. This suggests that other transcription factors account for the myeloid lineage bias in LT-HSCs[52] and activate the neutrophil lineage trajectory through binding the *Cebpa* +37kb enhancer in a subset of HSPCs [9,53]. Although Fluidigm C1 exhibits better sensitivity [54] and fewer dropout events [55,56] than droplet-based approaches, it remains possible that LT-HSCs with very low *Cebpa* levels (below the detection threshold of 0.25 TPM) exist. The physiological relevance of such low transcript levels, which could be attributable to pervasive transcription, is unknown. It is equally possible that the few *Cebpa*-expressing LT-HSCs identified are the result of technical noise or phenotypic misclassification.

In the enhancer-deleted models, LT-HSC loss was proportional to the degree of neutropenia, suggesting a causal relationship between them. Although an alternative explanation could be that LT-HSCs are less sensitive to *Cebpa* levels than myeloid progenitors, the normal production of lymphoid cells indicates that LT-HSCs remain viable in the absence of *Cebpa*. This is further supported by the almost complete absence of *Cebpa*-expressing LT-HSCs in healthy bone marrow. Critically, LT-HSCs were also lost in mice transplanted with +37 kb^{HOM} bone marrow: 40% of the transplanted mice showed this severe phenotype, while the other 60% showed signs of myeloid differentiation defects and dysplasia, suggesting they were in an early stage of the same pathological process. These effects we can only explain by

systemic consequences triggered in the presence of neutropenia. Even in a hybrid model in which a few LT-HSCs express *Cebpa*, the evidence gathered in this study supports the hypothesis that the major LT-HSC depletion in *Cebpa* null mice is an indirect consequence of neutropenia.

To our knowledge, none of previously reported neutropenia mouse models[26,57] demonstrated a depletion of LT-HSCs. A likely explanation is that the differentiation block in these other models (such as Sbds [57] and Gfi1 [58] mutants) often occurs at a late stage of neutrophil differentiation. Niches supporting a late stage of neutrophil differentiation are located distant from LT-HSC niches. Given that the differentiation block in *Cebpa* null bone marrow occurs in early progenitors located proximally to LT-HSC niches, the proposed systemic effect of neutropenia onto the LT-HSCs might be specific to this model. Fifteen percent of patients with congenital neutropenia develop LT-HSC clonal bone marrow conditions such as myelodysplasia and AML [59,60], suggesting LT-HSC impairment may also occur in human neutropenia. Thus other models of neutropenia are required in order to understand the mechanisms that lead to hematopoietic insufficiency in the presence of neutropenia, such as *Cebpa*-+37kb-Enh(f/f);Mx1-Cre mice [12]. The advantage of those mice with an inducible +37kb enhancer-deletion system is that one could avoid enhancer-deletion during embryogenesis.

Loss of HSC quiescence is one of the hallmarks identified at the transcriptional level in the +37kb^{HOM} enhancer-deleted mice. The activation of compensatory mechanisms forcing neutrophil differentiation in myeloid biased progenitors is a potential underlying cause. For example, increased production of granulocyte colony stimulating factor (G-CSF) in the absence of peripheral neutrophils is seen in patients with neutropenia [61,62]. G-CSF stimulates granulopoiesis through binding to G-CSF receptor-expressing progenitors [63] and activates HSCs through attenuation of the Cxcr4-Cxcl12 retention factors expressed on HSCs and bone marrow stromal cells, respectively [64,65]. Therefore, the neutropenia-GCSF-HSC activation loop can eventually lead to exhaustion and consumption of the LT-HSC pool. Other potential causes that might lead to LT-HSC quiescence exit include metabolic stress caused by impaired differentiation [66]; emergency myelopoiesis due to infections acquired in the absence of neutrophils [67]; or the lack of mature neutrophils in the bone marrow niche that support HSC quiescence [68].

Myeloid niches in the bone marrow have been reported to be located spatially distant from niches occupied by HSCs [69] or lymphoid progenitor populations [70]. In addition, the dendritic, neutrophil and monocyte lineages are also distantly located from each other and organized in different sinusoid niches. Therefore, it is likely that the neutrophil-primed progenitors derived from both donor and recipient share a common environment that is potentially disturbed upon transplantation of +37kb^{HOM} bone marrow. The question that remains to be answered is how transplanted +37kb^{HOM} myeloid progenitors impair the differentiation process of the host. Transcriptome analysis shows that myeloid progenitors (cKit+ Sca1- Cd34+) derived from the +37kb enhancer-deleted model are metabolically

reprogrammed and exhibit downregulation of the oxidative phosphorylation (OXPHOS) pathway (Figure S8a, Supplementary Table 6), as compared to normal myeloid progenitors that are dependent on oxidative and mitochondrial metabolism [71]. Shutting down OXPHOS eventually activates the glycolytic pathway (known as the Warburg effect), under the control of nuclear factors such as *Mycn* [72,73], a critical metabolic regulator involved in cell competition. In line with this, we find *Mycn* expressed at high levels in the +37kb^{HOM} myeloid progenitors (Figure S8b). Such metabolic changes impact cell-to-cell communication processes required for normal differentiation programs [66], which may partially explain the acquired neutropenia in the recipient when transplanted with +37kb^{HOM} bone marrows.

Our study suggests that prolonged neutropenia induced perturbations in localized myeloid niches and further caused systemic bone marrow changes resulting in LT-HSC loss, bone marrow hypocellularity and severe dysplasia. Although the underlying mechanisms remain unclear, we hypothesize that *Cebpa* null progenitors acquire metabolic reprogramming that impairs differentiation and disturbs HSC quiescence. Functional studies are required to investigate the intracellular role of *Cebpa* in controlling the metabolic pathways related to neutrophil differentiation, and further elucidate how *Cebpa* null progenitors are metabolically reprogrammed to inflict systemic changes in the bone marrow. From a clinical perspective, this phenomenon may explain how metabolic stress on HSCs might give rise to bone marrow clonal disorders in a subset of congenital neutropenia patients. Therefore, our study sets a paradigm about the underlying mechanisms involved in the progression of neutropenia to clonal bone marrow disorders.

AUTHOR CONTRIBUTIONS

R.A., R.M-L and RD designed the research and wrote the paper. R.A, M.H, L.S, N.S., E.B performed Research, R.A, R.M-L, R.H, N.S, R.K.S, L.G analysed data.

My contributions to this work were: analysis of both bulk and single-cell RNA-seq data; interpretation of the results; and preparation of the manuscript.

ACKNOWLEDGEMENTS

We are grateful to Michael Vermeulen for the flowcytometric analysis and to our colleagues from the Erasmus Animal Facility who took care of the mice used in this study and conducted regular bleeding to obtain blood cells. We would also like to thank Stefan Erkeland, Zhen Ping and Jacqueline Feyen for their valuable input. The work was funded by grants and fellowships from the Dutch Cancer Society (EMCR 2015-7935) (R. D., R. A., R.M-L., M.H., L.S.) and the Lady Tata Foundation (R. A.).

CONFLICT OF INTEREST DISCLOSURE

None of the authors has a relevant conflict of interest.

REFERENCES

1. Becker AJ, McCulloch EA, Till JE. Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. *Nature*. 1963;197(4866):452–454.
2. Till JE, Mcculloch EA. A Direct Measurement of the Radiation Sensitivity of Normal Mouse Bone Marrow Cells. 1961.
3. Giladi A, Paul F, Herzog Y, et al. Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nat. Cell Biol.* 2018;20(7):836–846.
4. Rodriguez-Fraticelli AE, Wolock SL, Weinreb CS, et al. Clonal analysis of lineage fate in native haematopoiesis. *Nature*. 2018;553(7687):212–216.
5. Belluschi S, Calderbank EF, Ciaurro V, et al. Myelo-lymphoid lineage restriction occurs in the human haematopoietic stem cell compartment before lymphoid-primed multipotent progenitors. *Nat. Commun.* 2018;9(1):.
6. Weinreb C, Rodriguez-Fraticelli A, Camargo FD, Klein AM. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science (80-.).* 2020;367(6479):.
7. Moignard V, MacAulay IC, Swiers G, et al. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat. Cell Biol.* 2013;15(4):363–372.
8. Avellino R, Delwel R. Expression and regulation of C/EBPα in normal myelopoiesis and in malignant transformation. *Blood*. 2017;129(15):2083–2091.
9. Avellino R, Havermans M, Erpelinck C, et al. An autonomous CEBPA enhancer specific for myeloid-lineage priming and neutrophilic differentiation. *Blood*. 2016;127(24):2991–3003.
10. Cooper S, Guo H, Friedman AD. The +37 kb Cebpa enhancer is critical for Cebpa myeloid gene expression and contains functional sites that bind SCL, GATA2, C/EBPα, PU.1, and additional Ets factors. *PLoS One*. 2015;10(5):e0126385.
11. Hasemann MS, Lauridsen FKB, Waage J, et al. C/EBPα Is Required for Long-Term Self-Renewal and Lineage Priming of Hematopoietic Stem Cells and for the Maintenance of Epigenetic Configurations in Multipotent Progenitors. *PLoS Genet.* 2014;10(1):.
12. Guo H, Cooper S, Friedman AD. In vivo deletion of the Cebpa +37 kb enhancer markedly reduces Cebpa mRNA in myeloid progenitors but not in non- hematopoietic tissues to impair granulopoiesis. *PLoS One*. 2016;11(3):1–23.
13. Zhang P, Iwasaki-Arai J, Iwasaki H, et al. Enhancement of Hematopoietic Stem Cell Repopulating Capacity and Self-Renewal in the Absence of the Transcription Factor C/EBPα. *Immunity*. 2004;21(6):853–863.
14. Ye M, Zhang H, Amabile G, et al. C/EBPa controls acquisition and maintenance of adult haematopoietic stem cell quiescence. *Nat. Cell Biol.* 2013;15(4):385–394.
15. Busch K, Klapproth K, Barile M, et al. Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature*. 2015;518(7540):542–546.
16. Sun J, Ramos A, Chapman B, et al. Clonal dynamics of native haematopoiesis. *Nature*. 2014;514(7522):322–327.
17. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*. 2009;25(9):1105–11.
18. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 2010;28(5):511–515.

19. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016;44(D1):D733–D745.
20. Anders S, Pyl PT, Huber W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015;31(2):166–169.
21. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550.
22. Laurenti E, Frelin C, Xie S, et al. CDK6 levels regulate quiescence exit in human hematopoietic stem cells. *Cell Stem Cell*. 2015;16(3):302–313.
23. Fukushima T, Tanaka Y, Hamey FK, et al. Discrimination of Dormant and Active Hematopoietic Stem Cells by G0 Marker Reveals Dormancy Regulation by Cytoplasmic Calcium. *Cell Rep.* 2019;29(12):4144--4158.e7.
24. Cabezas-Wallscheid N, Klimmeck D, Hansson J, et al. Identification of regulatory networks in HSCs and their immediate progeny via integrated proteome, transcriptome, and DNA methylome analysis. *Cell Stem Cell*. 2014;15(4):507–522.
25. Yáñez A, Coetzee SG, Olsson A, et al. Granulocyte-Monocyte Progenitors and Monocyte-Dendritic Cell Progenitors Independently Produce Functionally Distinct Monocytes. *Immunity*. 2017;47(5):890-902.e4.
26. Muench DE, Olsson A, Ferchen K, et al. Mouse models of neutropenia reveal progenitor-stage-specific defects. *Nature*. 2020;(September 2018):
27. Hinge A, He J, Bartram J, et al. Asymmetrically Segregated Mitochondria Provide Cellular Memory of Hematopoietic Stem Cell Replicative History and Drive HSC Attrition. *Cell Stem Cell*. 2020;26(3):420-430.e6.
28. Olsson A, Venkatasubramanian M, Chaudhri VK, et al. Single-cell analysis of mixed-lineage states leading to a binary cell fate choice. *Nature*. 2016;537(7622):698–702.
29. Nestorowa S, Hamey FK, Pijuan Sala B, et al. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood*. 2016;128(8):e20–e31.
30. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* 2015;33(5):495–502.
31. Christodoulou C, Spencer JA, Yeh SCA, et al. Live-animal imaging of native haematopoietic stem and progenitor cells. *Nature*. 2020;578(7794):278–283.
32. Cabezas-Wallscheid N, Buettner F, Sommerkamp P, et al. Vitamin A-Retinoic Acid Signaling Regulates Hematopoietic Stem Cell Dormancy. *Cell*. 2017;169(5):807--823.e19.
33. Ivanova NB, Dimos JT, Schaniel C, et al. A stem cell molecular signature. *Science*. 2002;298(5593):601–4.
34. Eppert K, Takenaka K, Lechman ER, et al. Stem cell gene expression programs influence clinical outcome in human leukemia. *Nat. Med.* 2011;17(9):1086–1094.
35. Rodriguez-Fraticelli AE, Weinreb C, Wang SW, et al. Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis. *Nature*. 2020;583(June 2019):
36. Ni F, Yu WM, Wang X, et al. Ptpn21 Controls Hematopoietic Stem Cell Homeostasis and Biomechanics. *Cell Stem Cell*. 2019;24(4):608--620.e6.
37. Sugiyama T, Kohara H, Noda M, Nagasawa T. Maintenance of the Hematopoietic Stem Cell Pool by CXCL12-CXCR4 Chemokine Signaling in Bone Marrow Stromal Cell Niches. *Immunity*. 2006;25(6):977–988.
38. Lin KK, Rossi L, Boles NC, et al. CD81 is essential for the re-entry of hematopoietic stem cells to quiescence following stress-induced proliferation via deactivation of the Akt pathway. *PLoS Biol.* 2011;9(9):.
39. Zheng J, Umikawa M, Zhang S, et al. Ex vivo expanded hematopoietic stem cells overcome the MHC barrier in allogeneic transplantation. *Cell Stem Cell*. 2011;9(2):119–130.

40. Papathanasiou P, Attema JL, Karsunky H, et al. Evaluation of the long-term reconstituting subset of hematopoietic stem cells with CD150. *Stem Cells*. 2009;27(10):2498–2508.
41. Gatlin J, Melkus MW, Padgett A, Kelly PF, Garcia JV. Engraftment of NOD/SCID Mice with Human CD34 + Cells Transduced by Concentrated Oncoretroviral Vector Particles Pseudotyped with the Feline Endogenous Retrovirus (RD114) Envelope Protein. *J. Virol.* 2001;75(20):9995–9999.
42. Wiekmeijer A-S, Pike-Overzet K, Brugman MH, et al. Sustained Engraftment of Cryopreserved Human Bone Marrow CD34 + Cells in Young Adult NSG Mice. *Biores. Open Access*. 2014;3(3):110–116.
43. Zhang DE, Zhang P, Wang ND, et al. Absence of granulocyte colony-stimulating factor signaling and neutrophil development in CCAAT enhancer binding protein α-deficient mice. *Proc. Natl. Acad. Sci. U. S. A.* 1997;94(2):569–574.
44. Bereshchenko O, Mancini E, Moore S, et al. Hematopoietic Stem Cell Expansion Precedes the Generation of Committed Myeloid Leukemia-Initiating Cells in C/EBPα Mutant AML. *Cancer Cell*. 2009;16(5):390–400.
45. Kirstetter P, Schuster MB, Bereshchenko O, et al. Modeling of C/EBPα Mutant Acute Myeloid Leukemia Reveals a Common Expression Signature of Committed Myeloid Leukemia-Initiating Cells. *Cancer Cell*. 2008;13(4):299–310.
46. Pundhir S, Bratt Lauridsen FK, Schuster MB, et al. Enhancer and Transcription Factor Dynamics during Myeloid Differentiation Reveal an Early Differentiation Block in Cebpa null Progenitors. *Cell Rep.* 2018;23(9):2744–2757.
47. Pabst T, Mueller BU. Complexity of CEBPA dysregulation in human acute myeloid leukemia. *Clin. Cancer Res.* 2009;15(17):5303–5307.
48. Wouters BJ, Löwenberg B, Erpelinck-Verschueren CAJ, et al. Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome. *Blood*. 2009;113(13):3088–3091.
49. Stengel KR, Ellis JD, Spielman CL, Bomber ML, Hiebert SW. Definition of a small core transcriptional circuit regulated by AML1-ETO. *Mol. Cell*. 2021;81(3):530-545.e5.
50. Wilson M, Tsakraklides V, Tran M, et al. EVI1 interferes with myeloid maturation via transcriptional repression of Cebpa, via binding to two far downstream regulatory elements. *J. Biol. Chem.* 2016;291(26):13591–13607.
51. Jones LC, Lin ML, Chen SS, et al. Expression of C/EBPβ from the C/ebpα gene locus is sufficient for normal hematopoiesis in vivo. *Blood*. 2002;99(6):2032–2036.
52. Haas S, Trumpp A, Milsom MD. Causes and Consequences of Hematopoietic Stem Cell Heterogeneity. *Cell Stem Cell*. 2018;22(5):627–638.
53. Wilson NK, Foster SD, Wang X, et al. Combinatorial transcriptional control in blood stem/progenitor cells: Genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell*. 2010;7(4):532–544.
54. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med. 2017 91*. 2017;9(1):1–12.
55. Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun. 2018 91*. 2018;9(1):1–9.
56. Wang X, He Y, Zhang Q, Ren X, Zhang Z. Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2. *Genomics. Proteomics Bioinformatics*. 2021;
57. Zambetti NA, Bindels EMJ, Van Strien PMH, et al. Deficiency of the ribosome biogenesis gene Sbds in hematopoietic stem and progenitor cells causes neutropenia in mice by attenuating lineage progression in t al. Levels of human serum granulocyte colony-stimulating factor and granulocyte-macrophage colony-stimulating factor under pathological conditions. *Biotherapy*. 1992;4(2):147–153.
63. Tsuji K, Ebihara Y. Expression of G-CSF receptor on myeloid progenitors. *Leuk. Lymphoma*. 2001;42(6):1351–1357.

64. Broxmeyer HE, Orschell CM, Clapp DW, et al. Rapid mobilization of murine and human hematopoietic stem and progenitor cells with AMD3100, a CXCR4 antagonist. *J. Exp. Med.* 2005;201(8):1307–1318.

65. Karpova D, Ritchey JK, Holt MS, et al. Continuous blockade of CXCR4 results in dramatic mobilization and expansion of hematopoietic stem and progenitor cells. *Blood*. 2017;129(21):2939–2949.

66. Bracha AL, Ramanathan A, Huang S, Ingber DE, Schreiber SL. Carbon metabolism-mediated myogenic differentiation. *Nat. Chem. Biol.* 2010;6(3):202–204.

67. Manz MG, Boettcher S. Emergency granulopoiesis. *Nat. Rev. Immunol.* 2014;14(5):302–314.

68. Cossío I, Lucas D, Hidalgo A. Neutrophils as regulators of the hematopoietic niche. *Blood*. 2019;133(20):2140–2148.

69. Zhang J, Wu Q, Johnson CB, et al. In situ mapping identifies distinct vascular niches for myelopoiesis. *Nature*. 2021;590(7846):457–462.

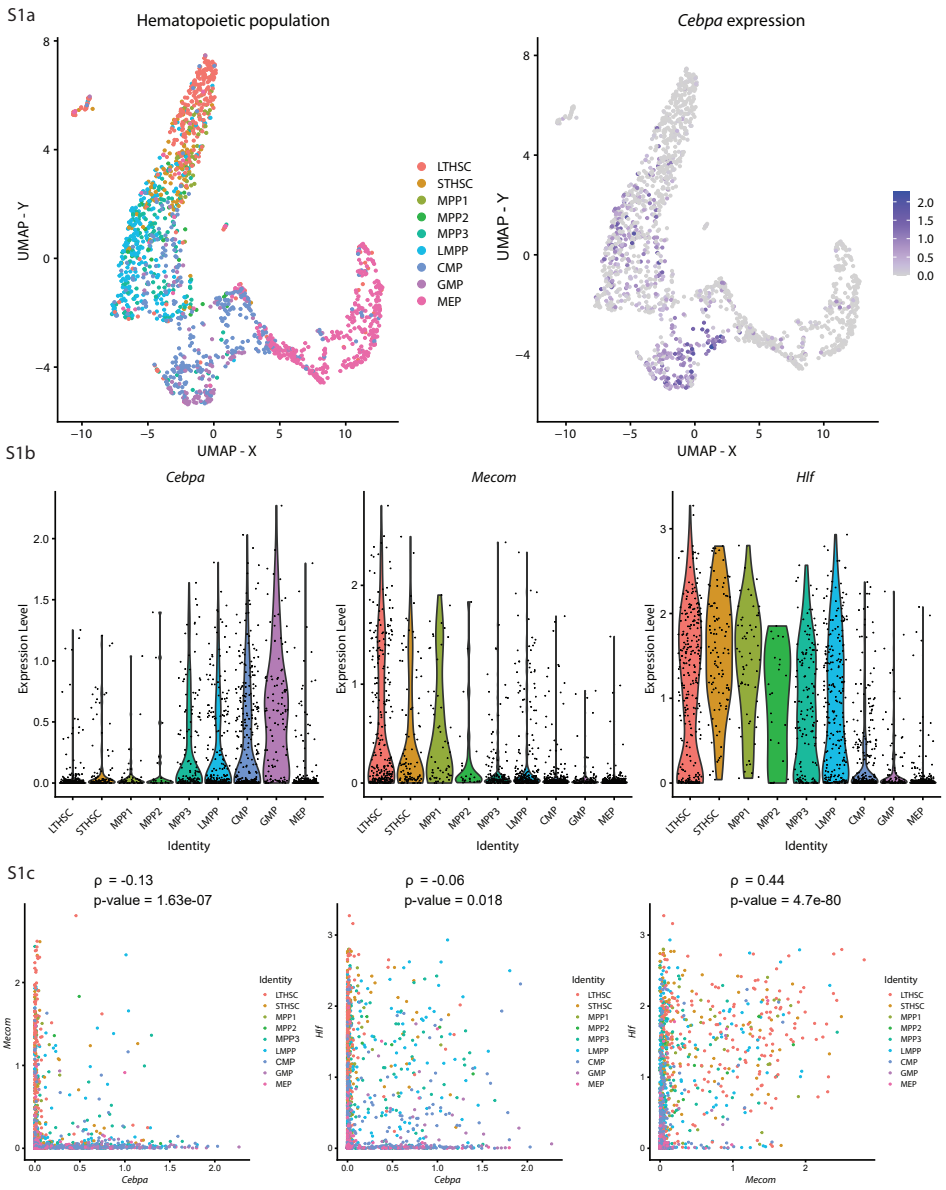
70. Cordeiro Gomes A, Hara T, Lim VY, et al. Hematopoietic Stem Cell Niches Produce Lineage-Instructive Signals to Control Multipotent Progenitor Differentiation. *Immunity*. 2016;45(6):1219–1231.

71. Laurenti E, Göttgens B. From haematopoietic stem cells to complex differentiation landscapes. *Nature*. 2018;553(7689):418–426.

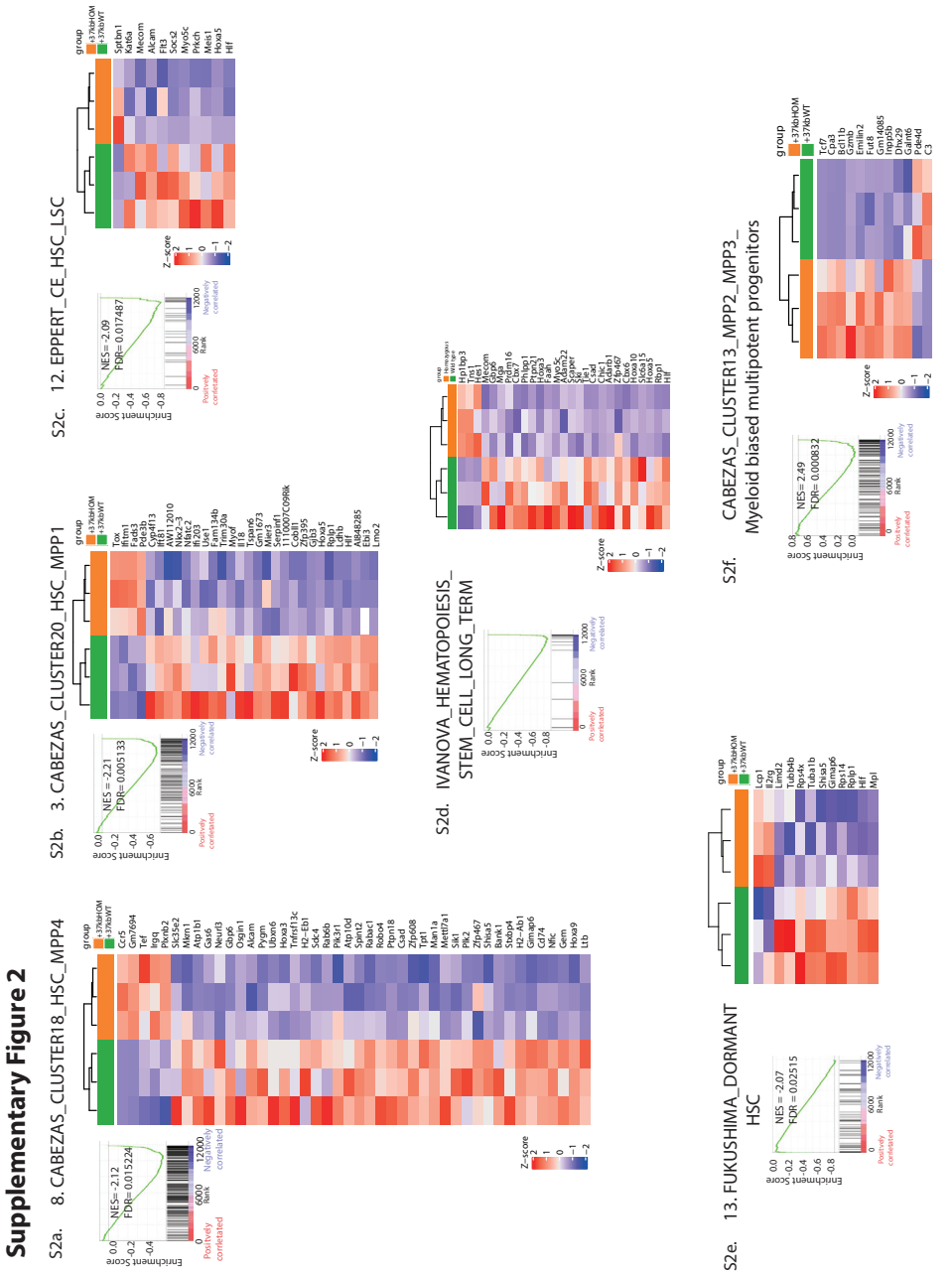
72. De La Cova C, Senoo-Matsuda N, Ziosi M, et al. Supercompetitor status of drosophila Myc cells requires p53 as a Fitness sensor to reprogram metabolism and promote viability. *Cell Metab.* 2014;19(3):470–483.

73. Tjaden B, Baum K, Marquardt V, et al. N-Myc-induced metabolic rewiring creates novel therapeutic vulnerabilities in neuroblastoma. *Sci. Rep.* 2020;10(1):1–10.

SUPPLEMENTARY FIGURES

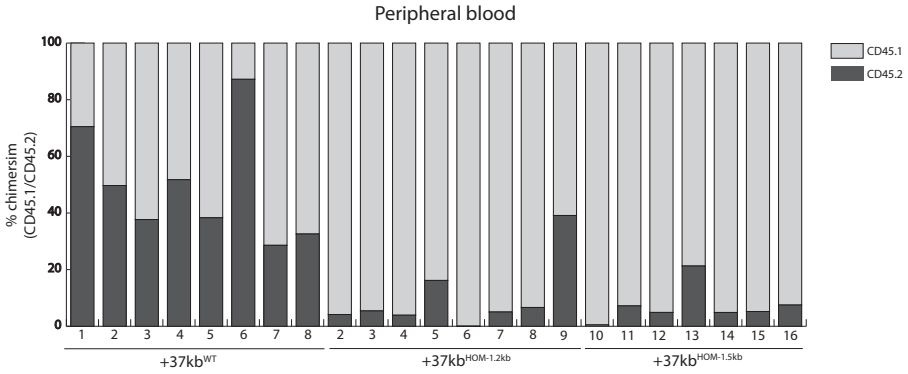


Supplementary Figure 1. Single cell RNA sequencing data from a separate cohort confirms that LT-HSCs rarely express *Cebpa*. (a) On the left, UMAP dimensionality reduction of scRNA-seq data from HSPCs annotated by index sorting. On the right, *Cebpa* expression projected on the UMAP. (b) Violin plots showing single cell expression of *Cebpa*, *Mecom* and *Hlf* in different HSPC subpopulations (c) Scatterplot showing that *Cebpa* and *Mecom* are largely mutually exclusive, whereas *Hlf* and *Cebpa* co-occur in some cells and *Mecom* and *Hlf* are frequently associated. Cells are color coded by the population they belong to. Pearson correlation coefficients and the related p-values for the pairwise gene combinations are depicted in the scatter plots. The data used in this figure were originally published by [27].

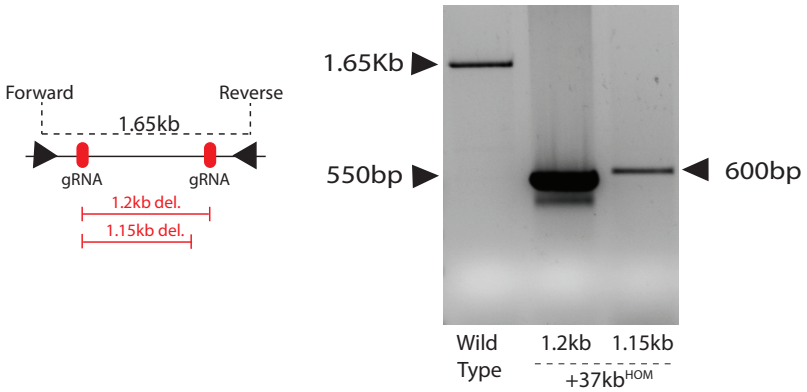


Supplementary Figure 2. **Transcriptome analysis of +37kb^{HOM} LSKs show loss of bone marrow quiescence in HSCPs.** For selected datasets derived from the literature, the GSEA enrichment plot is shown alongside a heatmap depicting the genes in that dataset with significant differential expression. The numbers preceding each cluster indicate the ranking in the GSEA, in decreasing order by normalized enrichment score. (a) Cluster #18, enriched in HSC and MPP4, from [22]. (b) Cluster #20, enriched in HSC and MPP1, from [22]. (c) Shared human HSC and acute myeloid leukemia stem cell (LSC) gene signature, from [32] (d) Genes upregulated in LT-HSCs from [31] (e) Genes upregulated in dormant HSCs compared to active HSCs, [21] (f) Cluster #18, enriched in MPP2 and MPP3, from [22].

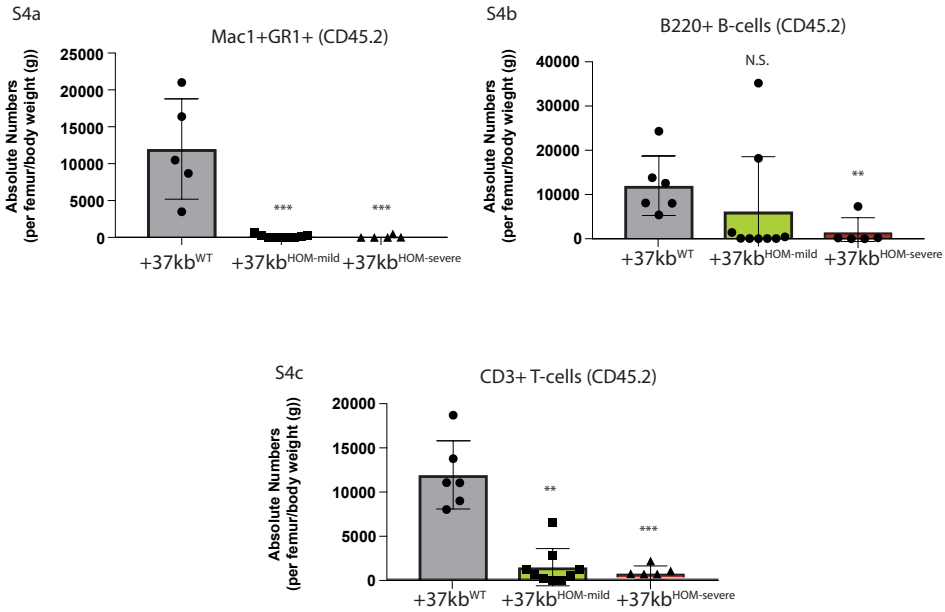
S3a



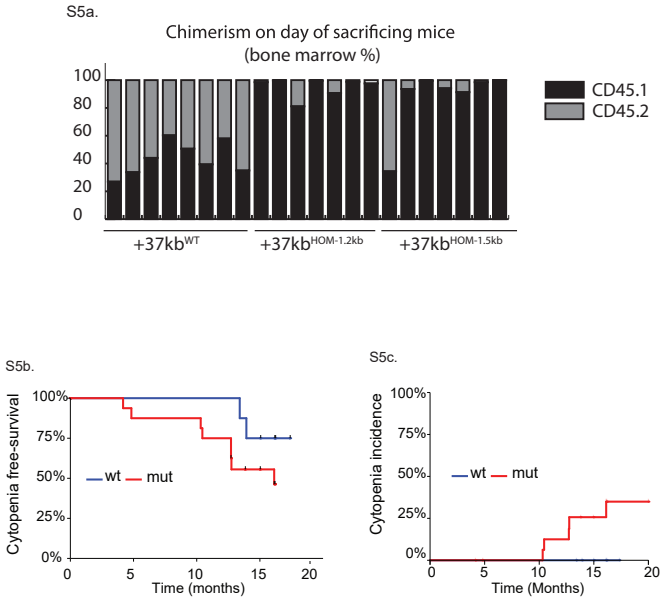
S3b



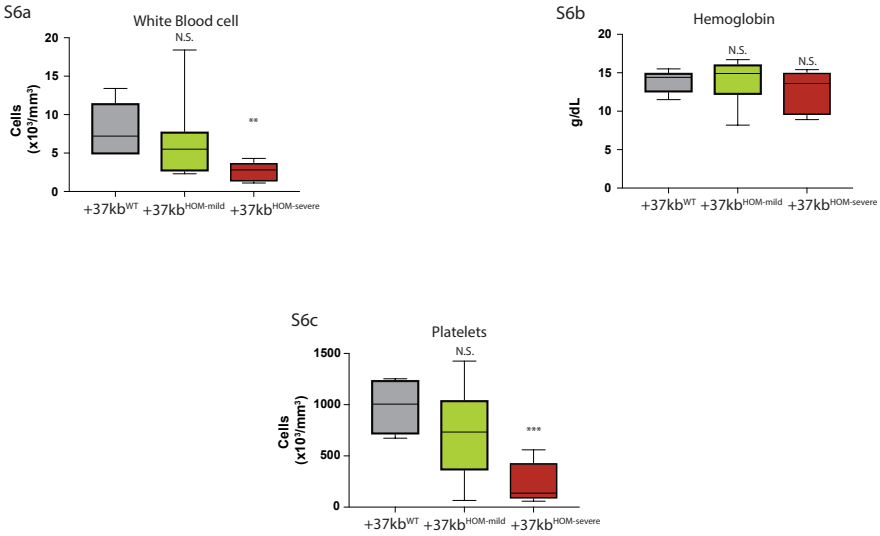
Supplementary Figure 3. (a) Bar chart showing percentage of CD45.1 and CD45.2 chimerism in peripheral blood, twelve weeks after transplantation. (b) (left) Schematic figure showing PCR setup to detect CRISPR targeted regions -- two guide RNAs (red) at the borders of the enhancer sequence, flanked by a pair of PCR primers (black triangles labelled forward and reverse). (right) PCR on genomic DNA from +37kb^{WT} and +37kb^{HOM} mice.



Supplementary Figure 4. Absolute numbers of (a) Mac1+Gr1+ myeloid cells, (b) B220+ B-cells and (c) CD3+ T-cells derived from bone marrows of CD45.2 +37kb^{WT} and +37kb^{HOM} (mild and severe) transplanted mice on the day they were euthanized. Cell numbers were calculated in a single femur and normalized by body weight (in grams) of each mouse analyzed.

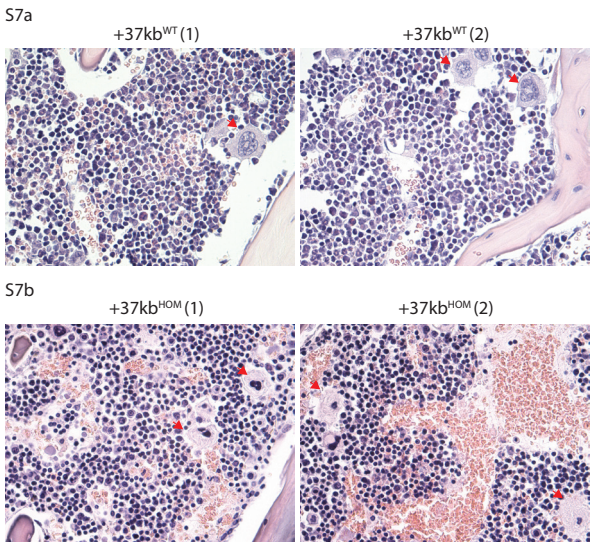


Supplementary Figure 5. (a) Percentage of chimerism between donor CD45.1 and CD45.2 mice on the day they were euthanized. (b) Cytopenia-free survival of +37kb^{HOM} mice compared to +37kb^{WT} (c) Incidence of neutropenia in +37kb^{HOM} mice compared to +37kb^{WT}



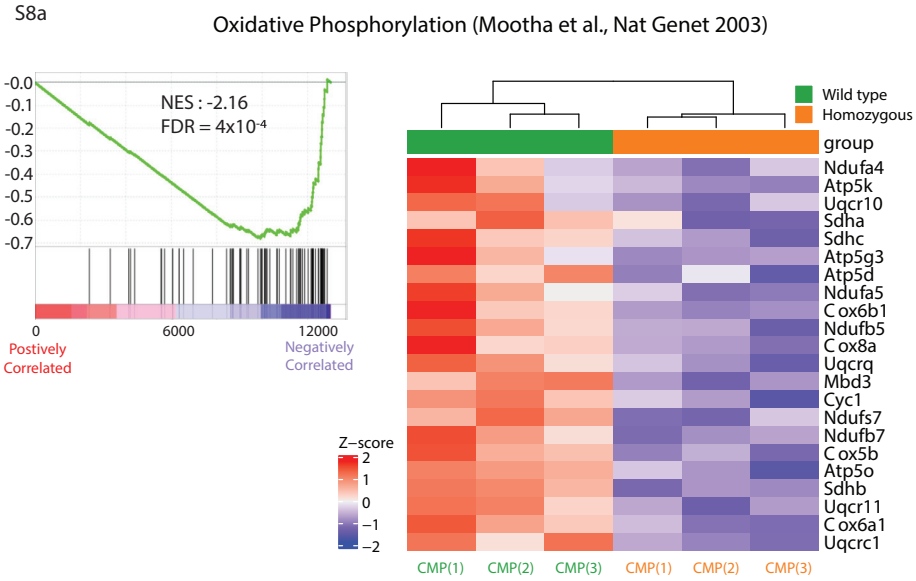
Supplementary Figure 6. Hematological parameters in mice transplanted with 37kb^{WT} and +37kb^{HOM} (mild and severe) bone marrow (a) White blood cell counts (b) Hemoglobin levels (c) Platelet counts

Supplementary Figure 7

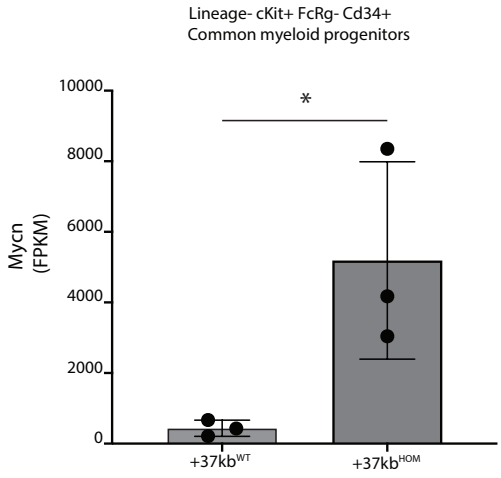


Supplementary Figure 7. Dysplastic megakaryocytes in +37kb^{HOM} mice. Hematoxylin and Eosin (H&E) staining of representative cross sections showing bone marrow architecture, identifying megakaryocytes with red arrows in +37kb^{WT} and +37kb^{HOM} mice.

S8a



S8b



Supplementary Figure 8. Metabolic reprogramming of myeloid progenitors. (a) Gene expression levels (FPKM) of *Mycn* in +37kb^{HOM} mice compared to +37kb^{WT}. *Mycn* is a critical metabolic regulator involved in cell competition. (b) On the left, GSEA enrichment plot showing downregulation of the gene set "Oxidative phosphorylation" (NES = -2.16, FDR < 0.05), derived from Mootha et al., Nat Genet 2003. On the right, heatmap of differentially expressed genes of this gene set in +37kb^{HOM} mice compared to +37kb^{WT}.

CHAPTER 3

Atypical 3q26/*MECOM* rearrangements genocopy inv(3)/t(3;3) in acute myeloid leukemia

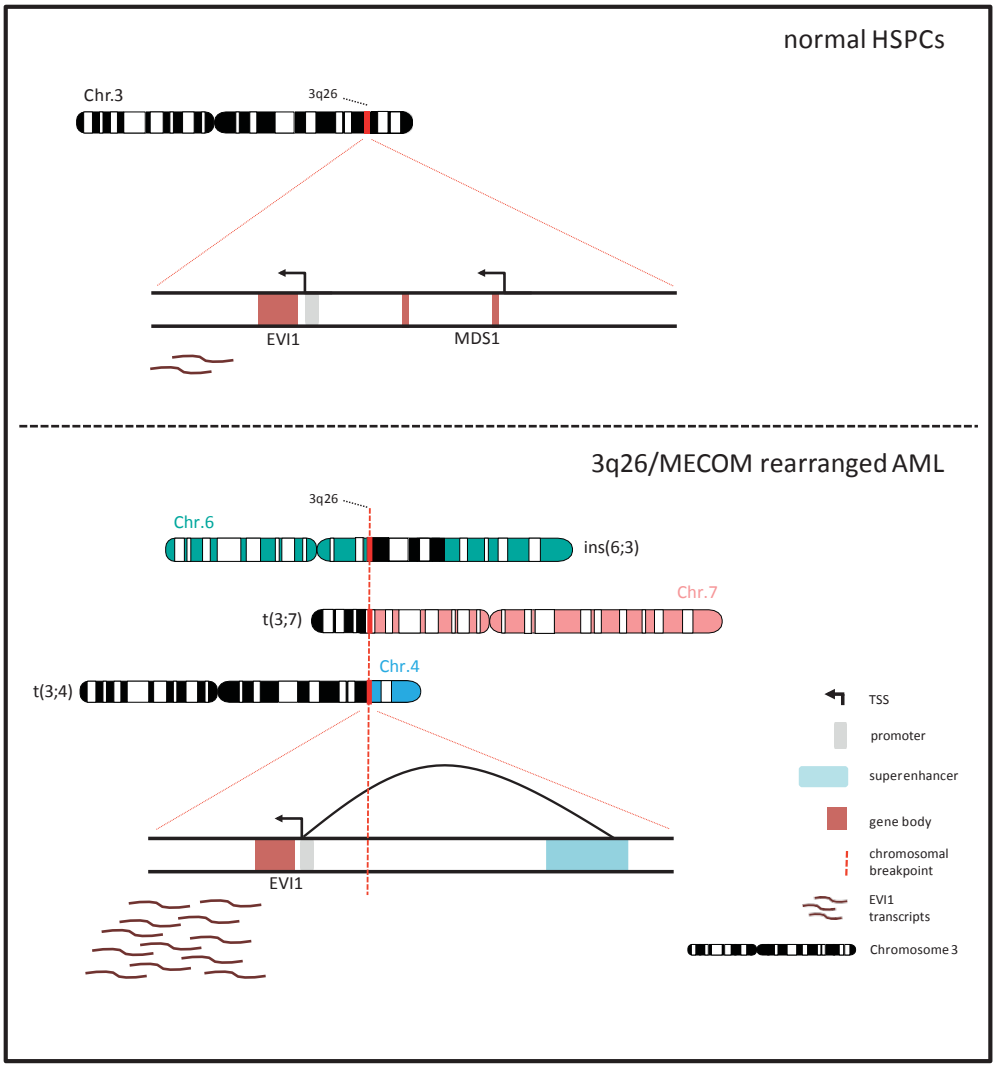
Sophie Ottema[1,2,*], **Roger Mulet-Lazaro[1,2,*],** H. Berna Beverloo[3], Claudia Erpelinck[1,2], Stanley van Herk[1,2], Robert van der Helm[3], Marije Havermans[1,2], Tim Grob[1], Peter J. M. Valk[1], Eric Bindels[1], Torsten Haferlach[4], Claudia Haferlach[4], Leonie Smeenk[1,2], Ruud Delwel[1,2]

[1] Department of Hematology, Erasmus University Medical Center, Rotterdam, The Netherlands
[2] Oncode Institute, Erasmus University Medical Center, Rotterdam, The Netherlands
[3] Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, The Netherlands
[4] MLL Munich Leukemia Laboratory, Munich, Germany.

*****These authors contributed equally to this work

Running Title: Oncogenic EVI1 upregulation in 3q26-rearranged AML

Blood. 2020 Jul 9;136(2):224-234. doi: 10.1182/blood.2019003701. PMID: 32219447.


normal HSPCs
Chr.3
3q26
EVI1
MDS1
3q26/MECOM rearranged AML
3q26
Chr.6
ins(6;3)
Chr.7
t(3;7)
Chr.4
t(3;4)
EVI1
TSS
promoter
superenhancer
gene body
chromosomal breakpoint
EVI1 transcripts
Chromosome 3

ABSTRACT

Acute myeloid leukemia (AML) with inv(3)/t(3;3)(q21q26) is a distinct WHO recognized entity, characterized by its aggressive course and poor prognosis. In this subtype of AML, the translocation of a *GATA2* enhancer (3q21) to *MECOM* (3q26) results in overexpression of the *MECOM* isoform *EVI1* and monoallelic expression of *GATA2* from the unaffected allele. The full-length *MECOM* transcript, *MDS1-EVI1*, is not expressed as the result of the 3q26 rearrangement. Besides the classical inv(3)/t(3;3), a number of other 3q26/*MECOM* rearrangements with poor treatment response have been reported in AML. Here we demonstrate, in a group of 33 AML patients with atypical 3q26 rearrangements, *MECOM* involvement with *EVI1* overexpression, but no or low *MDS1-EVI1* levels. Moreover, the 3q26 translocations in these AML patients often involve super-enhancers of genes active in myeloid development (e.g. *CD164*, *PROM1, CDK6 or MYC*). In more than 50% of these cases allele specific *GATA2* expression was observed, either by copy number loss or by an unexplained allelic imbalance. Altogether, atypical 3q26 recapitulate the main leukemic mechanism of inv(3)/t(3;3) AML, namely *EVI1* overexpression driven by enhancer hijacking, absent *MDS1-EVI1* expression and potential *GATA2* involvement. Therefore, we conclude that both atypical 3q26/*MECOM* and inv(3)/t(3;3) can be classified as a single entity of 3q26-rearranged AMLs. Routine analyses determining *MECOM* rearrangements, *EVI1* and *MDS1-EVI1* expression are required to recognize 3q-rearranged AML cases.

KEY POINTS

- *EVI1* overexpression, super-enhancer hijacking, lack of *MDS1-EVI1* and frequent *GATA2* **deficiency define 3q26/MECOM rearranged AML**
- 3q26/*MECOM* rearranged AML is a single entity, including but not limited to inv(3)/t(3;3), and requires specialized diagnostic assays

INTRODUCTION

Risk classification of patients with acute myeloid leukemia (AML) is based on the various genetic and epigenetic abnormalities previously identified and determines choice of treatment[1-5]. Understanding the biological consequences of these abnormalities is essential to develop new treatments for AML, especially for chemotherapy resistant subtypes. AML with inv(3)(q21q26) or t(3;3)(q21;q26)[6-9], henceforth referred to as inv(3)/t(3;3), is one of such subgroups with very poor response to therapy and a very aggressive course.

Recurrent translocations and inversions in AML most frequently generate oncogenic fusion genes[10-12]. However, in the case of an inv(3) or t(3;3), both rearrangements cause the translocation of an enhancer of the *GATA2* gene, located at 3q21, to the *MECOM* locus at chromosome 3q26[13,14]. *MECOM* encodes the transcript isoforms *MDS1-EVI1* and *EVI1*, which can be transcribed from two distinct promoters. In inv(3)/t(3;3) AML, the translocated *GATA2* enhancer causes overexpression *of EVI1*, but not of *MDS1-EVI1*. Translocation of the *GATA2* oncogenic enhancer in AML with an inv(3)/t(3;3) leads to *EVI1* upregulation and simultaneously abolishes *GATA2* expression from the rearranged allele[13,14]. Notably, germline haploinsufficiency or loss-of-function mutations in *GATA2* are the underlying causes of a wide spectrum of disorders, including MonoMAC and Emberger syndrome[15-18]. Those patients have a severely increased chance to develop AML compared to healthy individuals. Together with the fact that *GATA2* encodes a transcription factor essential for normal hematopoietic development[19], this suggests that loss of one *GATA2* allele increases the transforming ability of *EVI1* in chromosome 3q26-rearranged AMLs.

In a previous study of newly diagnosed 6515 AML patients, a group of leukemias with undefined 3q abnormalities was reported[9]. Although these patients did not present with a classical inv(3)/t(3;3), they also exhibited frequent *EVI1* overexpression and a very poor survival[9]. Here we addressed the question whether patients within this group harboring rearrangements at 3q26, resemble inv(3)/t(3;3) AML. Our study identifies critical similarities in the pathophysiology of both atypical 3q26 and inv(3)/t(3;3) AMLs: myeloid enhancer-driven *EVI1* overexpression, accompanied by low or no *MDS1-EVI1* transcription and, in approximately 50% of the cases *GATA2* deficiency. Given their clinical and biological similarities, we conclude that atypical 3q26-rearranged AML and inv(3)/t(3;3) constitute a single entity.

METHODS

Patient material

Samples of the selected patients presenting with MDS or AML were collected either from the Erasmus MC Hematology department biobank (Rotterdam, The Netherlands) or from the MLL Munich Leukemia Laboratory biobank (Munich, Germany). Leukemic blast cells were purified from bone marrow or blood by standard diagnostic procedures. All patients provided written informed consent in accordance with the Declaration of Helsinki.

Cytogenetics: karyotype and FISH

Diagnostic cytogenetics for all samples was performed by each of the institutes mentioned above. For this study, samples were selected based on 3q26 rearrangements (other than recurrent or classic 3q26 rearrangements) detected by karyotyping or *MECOM* interphase fluorescence *in situ* hybridization (FISH). FISH and classic metaphase karyotyping were performed and reported according to standard protocols based on the International System of Human Cytogenetics Nomenclature (ISCN) 2017[20]. *MECOM* FISH was performed according to the manufacturer's protocol, using the *MECOM* t(3;3); inv(3)(3q26) triple color probe (Cytocell, LPH-036).

RNA isolation and qPCR

RNA was isolated using phenol-chloroform extraction followed by DNase digestion or using the Qiagen Allprep DNA/RNA kit and protocol (Qiagen, #80204). cDNA synthesis was done using the SuperScript II Reverse Transcriptase kit (Invitrogen). Quantitative real-time PCR was performed by using primers as described previously[13,21] on the 7500 Fast Real-time PCR System (Applied Biosystems). Relative levels of gene expression were calculated using the Δ*Ct* method[7,8,22].

SNP-Array

Patient blasts were stored at -80°C in RLT+ buffer (Qiagen) and DNA was isolated using the AllPrep DNA/RNA mini kit (Qiagen, #80204). All SNP-arrays were performed at the Erasmus MC Department of Clinical Genetics (Rotterdam, The Netherlands) as previously described [23,24]. In summary, per sample, 50-200 ng DNA was used for a single Illumina Global Screening Array (GSAMD)(San Diego, CA, USA). The array profiles were analyzed with a 0.15 Mb resolution in UCSC (Human Mar. 2006 (NCBI36/hg18) Assembly) by using Genome Studio (Illumina) and different versions of Nexus Copy Number Software (BioDiscovery: versions 5.0 and higher (Hawthorne, CA, USA)).

Targeted chromosomal region 3q21.1-3q26.2 DNA sequencing (3q-capture)

DNA was isolated as mentioned above. 3q-capture DNA sequencing was performed as we

described previously [13]. In summary, genomic DNA was fragmented using the Covaris shearing device (Covaris), and sample libraries were assembled following the TruSeq DNA Sample Preparation Guide (Illumina). After ligation of adapters and an amplification step, target sequences of chromosomal regions 3q21.1-q26.2 were captured using custom in-solution oligonucleotide baits (Nimblegen SeqCap EZ Choice XL). The design of target sequences was based on the human genome assembly hg19: chr3q21.1:126036241-130672290 - chr3q26.2:157712147-175694147. Amplified captured sample libraries were paired-end sequenced (2x100 bp) on the HiSeq 2500 platform (Illumina) and aligned against the hg19 reference genome using the Burrows-Wheeler Aligner (BWA)[25]. Chromosomal breakpoints were determined using Breakdancer [26]. All chromosomal aberrations found using this program were visually confirmed in the Integrated Genome Viewer (IGV)[27].

RNA sequencing

Sample libraries were prepped using 500 ng of input RNA according to the KAPA RNA HyperPrep Kit with RiboErase (HMR) (Roche) using Unique Dual Index adapters (Integrated DNA Technologies, Inc.). Amplified sample libraries were paired-end sequenced (2x100 bp) on the Novaseq 6000 platform (Illumina) and aligned against the human genome (hg19) using STAR version 2.5.4b. A description of the quantification and differential expression analysis is provided in the Supplementary Material.

Exome sequencing

DNA was isolated as described above. The Genomic DNA Clean & Concentrator kit (ZYMO Research) was used to remove EDTA from the DNA samples. Sample libraries were prepared using 100 ng of input according to the KAPA HyperPlus Kit (Roche) using Unique Dual Index adapters (Integrated DNA Technologies, Inc.). Exomes were captured using the SeqCap EZ MedExome (Roche Nimblegen) according to SeqCap EZ HyperCap Library v1.0 Guide (Roche) with the xGen Universal blockers – TS Mix (Integrated DNA Technologies, Inc.). The amplified captured sample libraries were paired-end sequenced (2x100 bp) on the Novaseq 6000 platform (Illumina) and aligned to the hg19 reference genome using the Burrows-Wheeler Aligner (BWA)[25]. A description of the variant calling and allele expression analysis is provided in the Supplementary Material.

Whole-genome sequencing

DNA isolation, whole genome library preparation and sequencing was performed at the Munich Leukemia Laboratory (MLL, Munich, Germany). Sequencing was performed on the Novaseq 6000 platform (Illumina). The experimental procedures are detailed in a previous report by the MLL laboratory[28]. WGS data were aligned to the hg19 reference genome using the Burrows-Wheeler Aligner (BWA)[25].

RESULTS

Frequent *MECOM* rearrangements in atypical 3q26 AML

To study *MECOM* involvement we performed Fluorescent In Situ Hybridization (*MECOM*-FISH, Figure S1A) in 33 AML patient samples whose karyotypes do not harbor a classical inv(3)/t(3;3) but had rearrangements at 3q26. These cases were classified as atypical 3q26 rearranged AML (Table 1, Table S1). A rearranged FISH pattern was found in 25 cases, i.e. a part of the *MECOM* signal was found translocated from chromosome band 3q26 to another locus in the genome (Table 1, Figure S1B). SNP-array hybridizations revealed losses or gains on 3q26 or and/or partner loci in 7 of these 25 cases (Table 1, Table S1). In 12 of these 25 *MECOM* rearranged cases, no copy number gains (CNG) or losses (CNL) were found, which is in agreement with the existence of balanced translocations (Table 1). In the remaining 6 it was unclear whether rearrangements were balanced or not. In 4 of the total cohort of 33 cases (#HF-13, 14, 15, 16), FISH analysis suggested amplification of the 3q26/*MECOM* locus (Table 1), which was confirmed by SNP-array (Table 1). In 2/33 atypical 3q26 samples (#TG-04, TG-06) no clear *MECOM* rearrangements could be detected. Together, these results point to common *MECOM* involvement in AML with atypical 3q26 rearrangements.

High *EVI1* mRNA levels transcribed from one allele in atypical 3q26 AML

Routine diagnostic RT-PCR[8] (Table 1) showed *EVI1* overexpression in 30 out of 33 atypical 3q26 cases. RNA sequencing (n=26) revealed that on average, *EVI1* transcript levels were over 9 fold higher (p=3.00e09) in atypical 3q26 AML than in control non-3q26 AML (Figure 1A). To discriminate between the two *MECOM* alleles, we assessed single nucleotide variants (SNVs) in RNA-seq and 3q-capture data. We could identify informative heterozygous SNPs in the DNA of 15 patients out of 33 patients and demonstrated equal distribution of the two *EVI1* alleles (Figure 1B, left bar in red and blue). RNA-seq data demonstrated monoallelic *EVI1* mRNA expression in those 15 leukemia samples (Figure 1B, right bar in red), strongly suggesting that *EVI1* is only transcribed from the rearranged *MECOM* allele in atypical 3q26 AML.

Low *MDS1-EVI1* expression is a common feature of atypical 3q26 AML

Although two messenger RNAs can be transcribed from the *MECOM* locus, i.e. *MDS1-EVI1 (ME)* and *EVI1* (Figure S1D)[29,30], inv(3)/t(3;3) AMLs are *EVI1+/ME-*. Similarly, in 29 out of 33 atypical 3q26 AML samples *MDS1-EVI1* transcripts were absent or expressed at very low levels as reported for inv(3)/t(3;3) leukemias (Table 1 and Figure 1C).

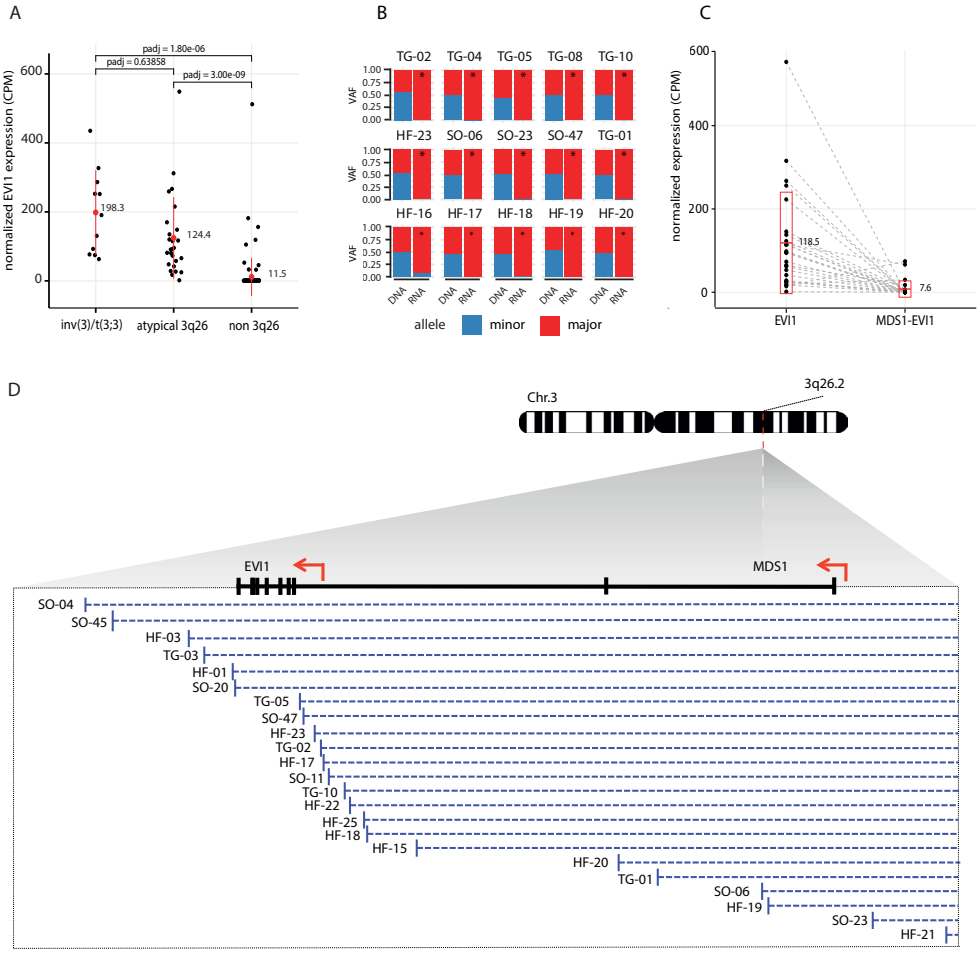


Figure 1. *MECOM* rearrangements, *EVI1* overexpression and absence of *MDS1-EVI1* expression in atypical 3q26 rearranged AML. (A) Normalized *EVI1* expression (counts per million (CPM) from RNA-seq data) determined in inv(3)/t(3;3) (N=11), atypical 3q26 (N=26) compared to non-3q26 AML (N=111). (B) Allele specific expression analysis using DNA-seq and RNA-seq data. The major allele is the allele of which the most SNPs were measured; the minor allele represents the allele that was underrepresented in the measurements. In order to perform this analysis, SNPs needed to be present in the sample. In 15/33 cases this analysis could be carried out. * indicates significant differential expression between alleles ($p<0.05$, χ^2 test). (C) Relative *EVI1* and *MDS1-EVI1* expression (CPM, RNA-seq) in atypical 3q26 AMLs (N=26). The red crossbar represents the mean and red box the standard deviation. (D) Schematic depiction of the breakpoints within the *MECOM* locus (3q26) determined by 3q-capture. The breakpoints could be determined in 23 AML cases. In 6 cases the breakpoint was 3'of *EVI1*, in 15 cases 5' of the *EVI1* promoter but 3' of the *MDS1-EVI1* promoter and in 2 AMLs 5'of the *MDS1-EVI1* promoter.

Frequent disruption of *MDS1* in atypical 3q26 AML underlies its low expression

In 23 out of 33 cases, we were able to exactly define the breakpoints within the *MECOM* locus (Figure 1D). Breakpoints occurred either upstream (N=17) or downstream (N=6) of the *EVI1* gene. In 15 out of the 17 cases with an upstream *EVI1* rearrangements, the breakpoints occurred between the MDS1 and EVI1 promoter (Figure 1D), as was reported in AML with a translocation t(3;3)(q21;q26)[15]. In those AMLs, the *MDS1* promoter has been dislocated due to the translocation, which avoids transcription of the long-form *MDS1-EVI1* (Figure S1D and 1C). In the 2 other AMLs (#-SO-23, HF-21) with a 5'-*EVI1* breakpoint, the rearrangements occurred upstream of the *MDS1* promoter. Accordingly, one of those patients (#SO-23) showed *EVI1+/ME+* expression. In the other case (#HF-21), neither *EVI1* nor *MDS1-EVI1* was detectable. The 6 cases with breakpoints 3' of *EVI1* showed an *EVI1+/ME-* expression pattern. Why 3q26 rearrangements with downstream breakpoints, as in AML with inv(3), show no or low *MDS1-EVI1* levels remains unresolved. CNV analysis of the 3q-capture DNA-seq and the SNP-array hybridizations revealed deletions within the *MDS1* region in 6 atypical 3q26 AML patients: #HF-15, HF-16, HF-20, HF21, TG-05, and SO-11 (Figure 2 and S3A, Table 1 and S1). Notably, these deletions underlie the loss of *MDS1* expression in #HF-16 and HF-21, where this cannot be explained by a translocation. *EVI1* exons were never deleted in those samples, and in fact were amplified in 3 of them (#HF-15, HF-16, TG-05). Altogether, the data strongly support the hypothesis that *EVI1* and not *MDS1-EVI1* expression is essential in transformation of 3q26-rearranged AMLs.

Unique rearrangements between *MECOM* and myeloid genes in atypical 3q26 AMLs

In 20/33 atypical 3q26 cases, the translocated partner locus of *MECOM*/3q26 could be identified by 3q-capture DNA-seq (Table 1). In two cases (#TG-03 and #SO-45) a cryptic inv(3)/t(3;3) *GATA2/MECOM* rearrangements was found. In 7 other cases, previously reported recurrent 3q26 translocations were identified, i.e. t(2;3)(p21;q26) (N=3), t(3;7)(q26;q21) (N=2), t(3;8)(q26;q24) (N=1) and t(3;6)(q26;q25) (N=1). The genes thought to be involved in those translocations are *THADA*, *CDK6*, MYC, and *ARID1B* respectively[31-36] (Table 1). These abnormalities were most probably missed at diagnosis due to the complex genetic nature of these cases. In the other 11 atypical 3q26 AMLs, novel and unique MECOM/3q26 rearranged partner loci were found (Table 1). We hypothesize that regulatory elements of these genes were hijacked by *EVI1*, resulting in loss of expression of the gene at the rearranged allele. Combined DNA-seq/RNA-seq SNP analysis applied to these AMLs revealed monoallelic or skewed expression of some of these genes in the translocated locus. As an example in AML with ins(6;3)(q21;q21q26) (#HF-23, Figure 3A), t(3;4)(q26;p15) (#HF-19, Figure 3B) or a t(3;7)(q26;p22) (#SO-20, Figure 3C) skewed expression of *CD164*, *PROM1* (*CD133*) or *FSCN1/EIF2AK1* were found respectively (Figure 3D). Whether the repressed allele was rearranged could not be assessed due to lack of patient material. These genes are all expressed in CD34+ cells and myeloid progenitor cells[37], and both *CD164*[38-40] and *PROM1*[41] are known to play a prominent role in hematopoiesis.

Table 1. Cytogenetic and *MECOM* associated alterations in atypical 3q26 AML

PT#	Karyotype Chr.3[1,2]	FISH EVI1[3]	SNP Chr.3[4]
SO-03	add(3)(q2?6)	Rearranged	Chr.3q26 balanced
SO-06	?der(3)(q2?)	Rearranged	Chr.3q26 CNL 5' MECOM
SO-11	der(3)add(3)(p1?2)add(3)(q2?6)	Rearranged	Chr.3q26 CNL 3' and 5' MECOM
SO-20	add(3)(q26)	Rearranged	Chr.3q26 balanced
SO-23	add(3)(q2?5)	Rearranged	Chr.3q26 balanced
SO-45	del(3)(q2?3q2?6)	Rearranged	Chr.3q26 balanced
SO-47	add(3)(q2?6)	Rearranged	Chr.3q26 balanced
BB-01	no 3q aberrations	Rearranged	Chr.3q26 CNL MDS1
TG-01	t(3;11)(q26;q2?4)	Rearranged	not done
TG-02	t(3;18)(q26;q1?)	Rearranged	not done
TG-03	no 3q aberrations	Rearranged	Chr.3q26 balanced
TG-04	ins(3;3)(q26;q21q26)	unclear	Chr.3q26 balanced
TG-05	?add(3)(q25)	Loss	Chr.3q26 CNL MDS, CNG EVI1
TG-06	add(3)(q26)	Normal	Chr.3q26 balanced
TG-08	-3[3],del(3)(q2?4)[7]	Loss	Chr.3q21 CNL GATA2
TG-10	-3	Rearranged	Chr.3q21 CNL GATA2
HF-01	der(7)t(3;7)(q26;q11.2)	Rearranged	not done
HF-02	der(7)t(3;7)(q26;q22)	Rearranged	not done
HF-03	der(7)t(3;7)(q26;q21)	Rearranged	not done
HF-04	der(7)t(3;7)(q26;p11)t(3;7)(q26;q21), -3	Rearranged	not done
HF-13	der(3)t(3;14)(q21;q?)	Amplified	Chr.3q26 CNG, MECOM balanced
HF-14	der(3)(::3p12->3q13::3q26->3q26::)	Amplified	Chr.3q26 CNG MECOM
HF-15	r(3)(p11q26)del(3)(q14q26)	Amplified	Chr.3q26 CNG EVI1/CNL MDS1, Chr.3q21 CNL GATA2
HF-16	der(2)ins(2;3)(q31;q22q26)	Amplified	Chr.q26 CNG 5' and 3' EVI1/CNL MDS1, Chr.3q21 CNL GATA2
HF-17	t(5;8)(p13;p21)	Rearranged	Chr.3q26 balanced
HF-18	t(2;3;6)(p15;q26;q26)	Rearranged	Chr.3q26 balanced
HF-19	t(3;4)(q26;p15)	Rearranged	Chr.3q26 balanced
HF-20	t(3;8)(q26;p23)	Rearranged	Chr.3q26 CNL MDS1
HF-21	der(8)t(3;8)(q26;p23)	Rearranged	Chr.3q26 CNL MDS1, Chr.3q21 CNL GATA2
HF-22	der(3)t(2;3)(p14;q26)	Rearranged	Chr.3q26 balanced
HF-23	ins(6;3)(q21;q21q26)	Rearranged	Chr.3q26 balanced
HF-24	der(3)del(3)(p12p26)inv(3)(p26q26)	Rearranged	Chr.3q26 balanced
HF-25	t(3;10)(q26;q21)	Rearranged	Chr.3q26 balanced

1. Cytogenetic aberrations with a specific focus on 3q26. Complete karyotype is provided in Table S1. **2.** Patient numbers BB-01, TG-03, TG-10 and HF-17 (#) did not show a 3q26 rearrangement by karyotyping, but were identified as rearranged by routine *MECOM* FISH. **3.** FISH was carried out as outlined in materials and methods section and scored as: normal, loss, amplified or rearranged. In sample TG-04 the FISH results were unclear. **4.** CNL: Copy Number Loss, CNG: Copy Number Gain. **5.** *EVI1+* and *MDS1-EVI1+* were determined as previously reported.[2-4] **6.** Partner gene: the gene(s), expressed in CD34+ cells, located in closest vicinity to the breakpoint is indicated. **7.** Fusion transcript.

EVI1[5]	MDS1-EVI1[5]	Breakpoint	Gene partner[6]
+	-	breakpoint not found	
+	-	inv(3;3)(p23q26), complex	TGFBR2
+	-	t(3;7)(q26;q11.23/q21.12), complex	DMTF1
+	-	t(3;7)(q26;p22.2), complex	TNRC18/FBXL18
+	+	t(3;6)(q26;q25)	ARID1B
+	-	t(3;3)(q21;q26)+t(3;16)(q26;q22.1), complex	GATA2
+	-	t(2;3)(p21;q26)	THADA
+	-	breakpoint not found	
+	+	t(3;11)(q26;q24)	HSPA8-MECOM[7]
+	-	t(3;18)(q26;q21)	MECOM-TCF4[7]
+	-	inv(3)(q21q26)	GATA2
+	+	breakpoint not found	
+	-	del(3)(q25.3-q26.2)	IL12A-AS1
+	-	breakpoint not found	
+	+	breakpoint not found	
+	-	t(3;6)(q26;p22)	TDP2/JARID2
+	-	t(3;7)(q26;q21)	CDK6
+	-	breakpoint not found	
+	-	t(3;7)(q26;q21)	CDK6
+	-	breakpoint not found	
+	-	breakpoint not found	
+	-	breakpoint not found	TRA2B-MECOM[7]
+	-	inv(3)(q13.33q26.2)	GTF2E1/STXBP5L
+	-	breakpoint not found	
+	-	t(3;8)(q26;q24.1)	MYC
+	-	t(2;3)(p21;q26) +t(3;5)(q26;q34) + t(3;6)(q26;q27)	THADA
+	-	t(3;4)(q26;p15)	PROM1, CD38
+	-	t(3;8)(q26;p23)	TNKS/MSRA
-	-	t(3;8)(q26;p24), complex	FAM135B
+	-	t(2;3)(p21;q26)	THADA
+	-	ins(6;3)(q21;q21q26)	CD164
-	-	breakpoint not found	
-	-	t(3;10)(q26;q21)	ARID5B

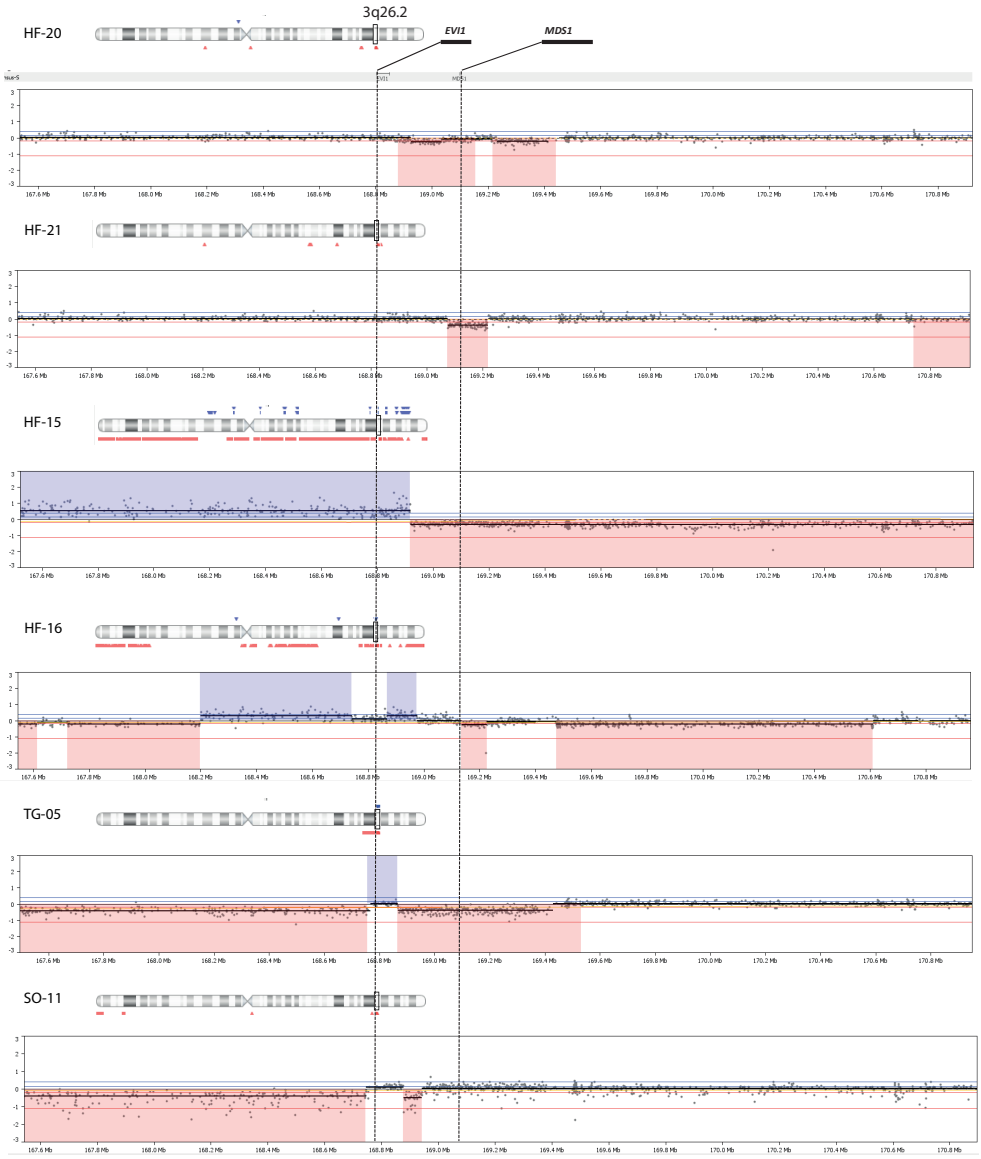


Figure 2. Copy number changes in the *MECOM* locus in atypical 3q26 AML. SNP array showing copy number losses (CNL) in red and copy number gains (CNG) in blue at chromosome band 3q26. *EVI1* and *MDS1-EVI1* are marked. Only the samples for which copy number changes were found in this locus are illustrated (N=6).

***MECOM* hijacks myeloid-specific enhancers that may activate *EVI1* transcription**

As chromatin of patient cells were not available, we studied the chromatin state at *CD164, PROM1* (*CD133*) and *FSCN1/EIF2AK1* in normal bone marrow CD34+ cells as well as in the inv(3) myeloid cell line MOLM-1[37,42]. As depicted in Figure 3A, 3B and 3C, binding of p300, presence of H3K27ac and lack of H3K4me3, were indicative of active enhancers within the regions that were translocated to *MECOM* in cases #HF-19, #HF-23 and #SO-20 respectively. In fact, the size of the H3K27 acetylated regions (>10kb) suggested the presence of a "super-enhancers"[43] in those loci (Figure 3E). Strong binding of key myeloid transcription factors like FLI1, GATA2 and RUNX1 (Figure 3A, 3B and 3C) in CD34+ bone marrow cells[37], further supports the notion that active myeloid "super-enhancers" translocate to *MECOM* in atypical 3q26 rearrangements to activate *EVI1* expression. ChIP-seq analysis of normal CD34+ and MOLM1 cells also showed the presence of "super-enhancers" in the regions near *THADA*, *MYC* and *CDK6*, that translocate to *MECOM* in AMLs with translocations t(2;3), t(3;8) and (t3;7) respectively (Table 1, Figure 3E, Figure S2A-E). The loss of these enhancers in one allele should lead to a reduction in total gene expression, but given that most of these translocations are unique to one patient, it is not possible to conduct a statistical analysis. Instead, for every gene that putatively loses its enhancer, we compared its average expression in the whole cohort to the expression in the individuals with the translocation. In line with our hypothesis, all genes except *MYC* exhibited reduced expression (Figure S3C). Together the data point to a mechanism of *EVI1* overexpression driven by hijacked myeloid "super-enhancers" in atypical 3q26 rearranged AML.

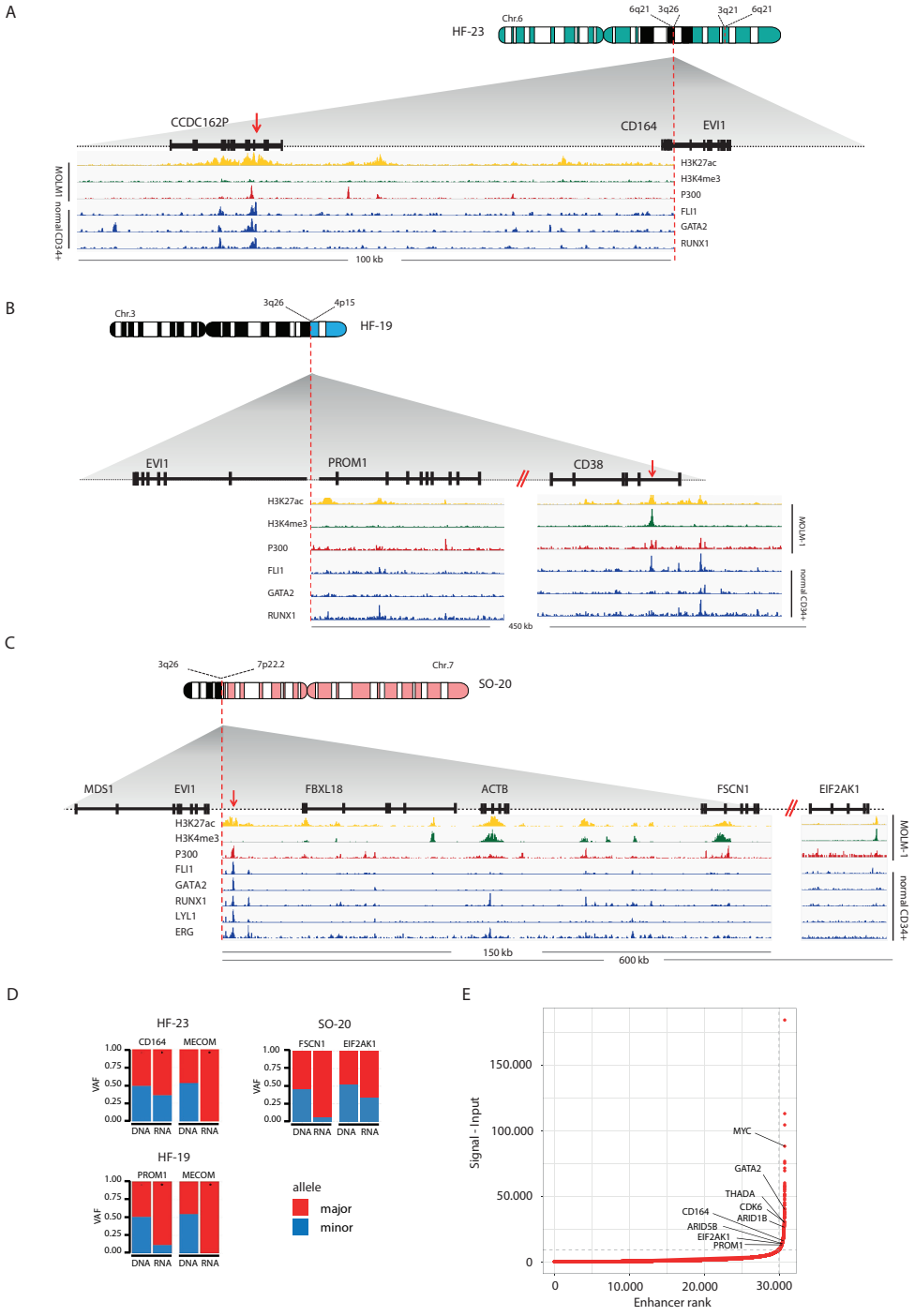

A
HF-23
Chr.6
6q21
3q26
3q21
6q21
CCDC162P
CD164
EVI1
H3K27ac
H3K4me3
P300
FLI1
GATA2
RUNX1
MOLM1
normal CD34+
100 kb
B
Chr.3
3q26
4p15
HF-19
EVI1
PROM1
CD38
H3K27ac
H3K4me3
P300
FLI1
GATA2
RUNX1
MOLM1
normal CD34+
450 kb
C
3q26
7p22.2
Chr.7
SO-20
MDS1
EVI1
FBXL18
ACTB
FSCN1
EIF2AK1
H3K27ac
H3K4me3
P300
FLI1
GATA2
RUNX1
LYL1
ERG
MOLM-1
normal CD34+
150 kb
600 kb
D
HF-23
CD164
MECOM
SO-20
FSCN1
EIF2AK1
HF-19
PROM1
MECOM
VAF
1.00
0.75
0.50
0.25
0.00
DNA RNA DNA RNA
allele
major
minor
E
Signal - Input
150.000
100.000
50.000
0
MYC
GATA2
THADA
CDK6
ARID1B
CD164
ARID5B
EIF2AK1
PROM1
0
10.000
20.000
30.000
Enhancer rank

Figure 3. Rearrangements involving *3q26/EVI1* and newly identified partner loci. (A), (B) and (C) Schematic depictions of chromosomal rearrangements of three unique atypical 3q26 patient samples, i.e. ins(6;3)(q26;q21q26) in patient #HF-23, t(3;4)(q26;p15) in patient #HF-19 and t(3;7)(q26;p22) in patient #SO-20 respectively. Figures show the loci and genes that have been rearranged and brought into the vicinity of *MECOM*: loci with *CD164* and *CCDC162P* (6q21) in A, *PROM1* and *CD38* (4p15) in B and *FBXL18, ACTB*, *FSCN1* and *EIF2AK1* (7p22) in C respectively. ChIP-seq tracks indicative for active enhancer elements, i.e. H3K27ac (yellow), H3K4me3 absence (green) and P300 (red), have been obtained from the MOLM-1 myeloid cell line[13]. Previously published ChIP-seq tracks of myeloid transcription factors FLI1, GATA2, RUNX1, LYL1 and ERG using normal CD34+ cells are shown[37] (blue). Enhancers possibly involved in *EVI1* activation are indicated with a red arrow. **(D)** Bar plots showing skewed expression of genes that putatively donated their enhancer. The bar plots show the genes with skewed expression, *CD164* (#HF-23), *PROM1* (#HF-19), *FSCN1* and *EIF2AK1* (#SO-20). In 2 out of 3 samples, monoallelic *EVI1* expression was found (#HF-23, #HF-19). Allele specific *EVI1* expression could not be determined in for #SO-20, since no SNPs could be detected. * Indicates significant differential expression between alleles ($p<0.05$, χ^2 test). **(E)** Hockey stick plot showing the classification of these long stretches of H3K27ac (A, B and C) found in the partner loci as super-enhancers (based on MOLM-1 H3K27ac ChIP-seq data using the ROSE algorithm).

Atypical 3q26 AMLs exhibit *GATA2* deficiency in half of the cases

In inv(3)/t(3;3) AML, the dislocation of the *GATA2* enhancer causes loss of expression of *GATA2* from the rearranged allele[13,14]. We addressed the question whether *GATA2* expression was reduced in atypical 3q26 AML without 3q21/*GATA2* rearrangements. RNA-seq data demonstrated comparable *GATA2* expression levels for the atypical 3q26 AMLs as for the inv(3)/t(3;3) AMLs (Figure 4A), which was slightly lower than in non-3q26 rearranged AMLs, although not statistically significant. Analysis of SNP-array data (performed for 27 atypical 3q26 AMLs) revealed copy number loss of parts of chromosome 3 including *GATA2* and/or its enhancer in 5 atypical 3q26 AML patients (#TG-08, TG-10, HF-15, HF-16 and HF-21, Figure 4C). In 2 of these cases loss of one chromosome 3 was also noted cytogenetically (Table 1). CNV analysis of the 3q-captured data of all 33 cases was used to verify copy number changes detected by SNP-array: 5 cases with *GATA2* or *GATA2*-enhancer loss were identified (Table S1) of which two are shown in Figure S3B. In 16 AMLs of our cohort, we could discriminate between two *GATA2* alleles based on SNP differences, identified by combined RNA- and DNA-seq data analysis. In 4 of those 16 cases *GATA2* expression was monoallelic or significantly skewed to one allele ($p<0.05$, marked by * in Figure 4B). As methylation of the *GATA2* promoters could explain allele specific expression, bisulfite-sequencing experiments were performed. However, we did not obtain any evidence for *GATA2* promoter methylation in these patients. Thus, the mechanism by which these cases showed unbalanced allelic *GATA2* expression remains unclear. Overall, we observed *GATA2* loss or skewed expression in 12 of the 22 (>50%) cases that we could analyze in full. No mutations in *GATA2* were found in any of the 33 atypical 3q26 AMLs. We conclude that in a subset of atypical 3q26 rearranged AML *EVI1* overexpression was accompanied by loss or diminished *GATA2* transcription from one allele, which resembles inv(3)/t(3;3) AML[13].

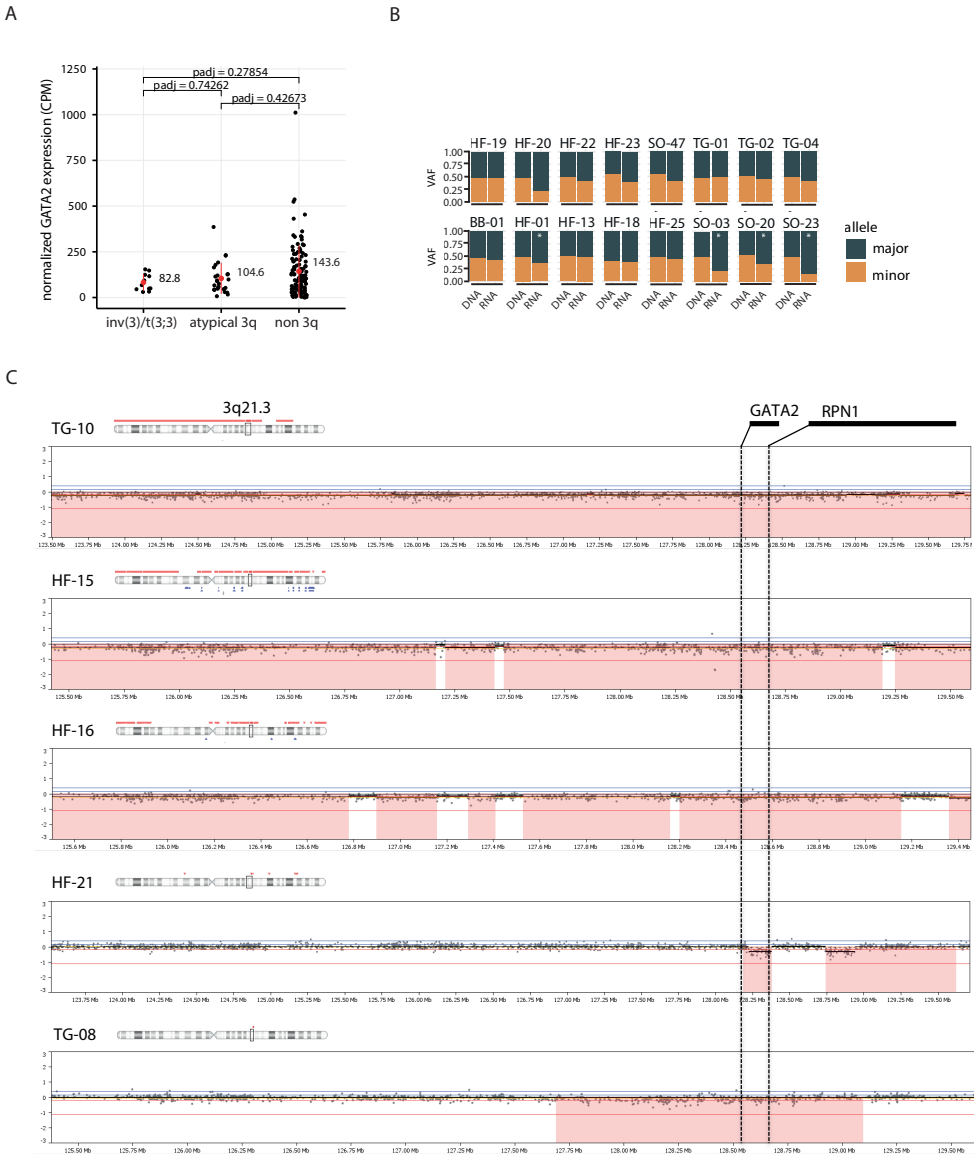


Figure 4. Copy number loss of *GATA2* or imbalanced *GATA2* expression in atypical 3q26 AML (A) *GATA2* expression (CMP, RNA-seq) determined in inv(3)/t(3;3) (N=11), in atypical 3q26 (N=26) and non-3q26 rearranged AML (N=111). Differences were not statistically significant (padj<0.05). Red dot represents the mean and the red bar the standard deviation. (B) Allele specific analysis using DNA-seq and RNA-seq data showed significant skewed expression of *GATA2* to one allele in 5 cases. In #HF-20 read depth was too low for a significance call. * indicates significant differential expression between alleles ($p<0.05$, χ^2 test). (C) SNP array data presented at chromosomal locus 3q21.3 using, showing CNLs in the *GATA2* locus, resulting in loss (red) of the *GATA2* gene or its enhancer (located in between *GATA2* and *RPN1*).

DISCUSSION

Atypical 3q26-rearranged AML represents a group of very poor risk leukemias with various undefined 3q26 rearrangements whose role in leukemogenesis is unclear[9]. Using a multipronged approach, we here demonstrate that in atypical 3q26-rearranged AML, *MECOM* is relocated, leading to *EVI1* overexpression in the absence of *MDS1-EVI1* transcription. We found potential myeloid super-enhancers to be translocated to *MECOM*. In approximately 50% of the study cohort *GATA2* skewed expression or copy number loss was found, despite lack of *GATA2* involvement in the rearrangement. We conclude that atypical 3q26 AML genocopy inv(3)/t(3;3) leukemias[13,14] and these two groups should be classified and treated as single entity.

In atypical 3q26 AMLs, chromosomal rearrangements bring *MECOM* into the vicinity of regulatory elements of genes active in myeloid cells, such as *THADA*, *CDK6, MYC*, *ARID1B, CD164*, *PROM1 (CD133) or FSCN1/EIF2AK1*[31-36]. We hypothesize that a mechanism of super-enhancer hijacking causes *EVI1* overexpression in variant 3q26-AMLs, as has been reported for the -77 kb *GATA2* enhancer in inv(3)/t(3;3) leukemias. ChIP-seq data from normal CD34+ bone marrow cells and myeloid cell lines revealed that transcription factors (TFs) that bind to the *GATA2* distal enhancer, including RUNX1, LYL1, SCL, FLI1, ERG, LMO2, and GATA2 itself[37], also interact with the loci translocated in atypical 3q26 AMLs. It will be challenging to model these translocations and study *EVI1* promoter interaction and regulation by these distinct super-enhancers. As super-enhancers have been reported to be hypersensitive to bromodomain-inhibitors[44,45], it will be interesting to study responses of the distinct 3q26-rearranged AMLs to those compounds.

It is well established that *EVI1* is an oncogenic driver of AML, but the role of *MDS1-EVI1* in leukemic transformation has not been thoroughly studied. *Evi1* was first identified as the ecotropic viral insertion site-1 in mouse leukemias, in which *Evi1* but not *Mds1-Evi1* was overexpressed due to retroviral insertions[46]. Patients with X-linked chronic granulomatous disease who received gene therapy to correct *GP91* (*PHOX*) mutations in hematopoietic progenitor cells, similarly developed AML due to retroviral insertions driving *EVI1* and not *MDS1-EVI1* overexpression[47]. Here we demonstrate that in atypical 3q26 AML, as reported in AML with inv(3)/t(3;3), overexpression of *EVI1* was accompanied by absence or low expression of *MDS1-EVI1*. We hypothesize that the translocated enhancers in 3q26-rearranged AMLs are able to contact and co-activate the promoter of *EVI1*, but not the promoter of *MDS1-EVI1*.

Monoallelic expression of *GATA2* is another hallmark of inv(3)/t(3;3), caused by loss of the *GATA2* enhancer at the rearranged allele. Does monoallelic *GATA2* play a role in leukemic transformation in inv(3)/t(3;3)? In over 50% of the atypical 3q26 AMLs analyzed, skewed or monoallelic expression of *GATA2* was evident, due to cryptic *GATA2/MECOM* translocation, deletion of *GATA2* or a regulatory element or by currently unknown mechanisms. *EVI1* overexpressing mice develop myeloid leukemias with a shorter latency when they are *GATA2*

heterozygous[48]. Moreover, individuals with inherited *GATA2* mutations or loss of expression of one allele have a high chance to develop AML[15-19]. Altogether, loss of one *GATA2* allele appears to have an effect on leukemia development. A larger patient cohort is required to investigate whether *GATA2* monoallelic expression has an impact on prognosis of 3q26-rearranged AML.

Atypical 3q26 AMLs are difficult to define, as they are cytogenetically complex and heterogeneous. This underscores the importance of routine molecular diagnostic assays to recognize this subgroup of AML patients. We propose to identify 3q26/MECOM rearrangements by using *MECOM* FISH (Figure S1), which is applied routinely in AML diagnostics. Quantitative *EVI1* and *MDS1-EVI1* mRNA expression analysis can be indicative for *EVI1* deregulation by enhancer hijacking. Together, this combined analysis can be used to classify this subgroup of AML patients.

AUTHOR CONTRIBUTIONS

S.O., R.M., R.D. designed the study. S.O., C.E., S.H., R.H., M.H., T.G. E.B. L.S. carried out experiments. R.M., S.O., H.B.B., L.S., C.H., T.H., P.J.M.V. analyzed data. H.B.B., P.J.M.V., T.H., C.H. provided samples and/or data. R.D., R.M., S.O. wrote the manuscript.
My contributions to this work were: processing and analysis of all high throughput sequencing data (3q-capture, RNA-seq, WES, WGS); interpretation of the results; and writing of the manuscript.

ACKNOWLEDGEMENTS

The authors are indebted to the colleagues from the bone marrow transplantation group and the molecular diagnostics laboratory of the department of Hematology at the Erasmus University Medical Center for storage of samples and molecular analysis of the leukemia cells (H. B. Beverloo, K. Joode, M. Wattel, R. M. van der Helm and P. J. M. Valk). For a part of the patient material and sequencing data the authors are thankful to the MLL Münchner Leukämielabor GmbH in Germany (C. Haferlach and T. Haferlach). We also thank Pieter Sonneveld and our colleagues of the Hematology department for their input, and especially Bas Wouters for critically reading the manuscript. This work was funded by grants and fellowships from the Dutch Cancer Society, “Koningin Wilhelmina Fonds” (R. Delwel, R. Mulet-Lazaro, S. Ottema, T. Grob), Skyline DX (S. Ottema) and the Daniel den Hoed Foundation (L. Smeenk).

CONFLICT OF INTEREST DISCLOSURE

T.H. and C.H.: MLL Munich Leukemia Laboratory: Employment, Equity Ownership. The other authors have nothing to disclose.

REFERENCES

1. Döhner H, Estey EH, Amadori S, et al. Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood*. 2010;115(3):453-474.
2. Papaemmanuil E, Gerstung M, Bullinger L, et al. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *New England Journal of Medicine*. 2016;374(23):2209-2221.
3. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *New England Journal of Medicine*. 2013;368(22):2059-2074.
4. Döhner H, Estey E, Grimwade D, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood*. 2017;129(4):424-447.
5. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood*. 2016;127(20):2391-2405.
6. Morishita K, Parganas E, William CL, et al. Activation of EVI1 gene expression in human acute myelogenous leukemias by translocations spanning 300-400 kilobases on chromosome band 3q26. *Proceedings of the National Academy of Sciences*. 1992;89(9):3937-3941.
7. Lugthart S, van Drunen E, van Norden Y, et al. High *EVI1* levels predict adverse outcome in acute myeloid leukemia: prevalence of EVI1 overexpression and chromosome 3q26 abnormalities underestimated. *Blood*. 2008;111(8):4329-4337.
8. Barjesteh van Waalwijk van Doorn-Khosrovani S, Erpelinck C, van Putten WLJ, et al. High EVI1 expression predicts poor survival in acute myeloid leukemia: a study of 319 de novo AML patients. *Blood*. 2003;101(3):837-845.
9. Lugthart S, Gröschel S, Beverloo HB, et al. Clinical, Molecular, and Prognostic Significance of WHO Type inv(3)(q21q26.2)/t(3;3)(q21;q26.2) and Various Other 3q Abnormalities in Acute Myeloid Leukemia. *Journal of Clinical Oncology*. 2010;28(24):3890-3898.
10. Mitelman F, Johansson B, Mertens F. Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nature Genetics*. 2004;36(4):331-334.
11. Mitelman F JBaMF. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer (2019). Nat. Genet., 36 (2004), pp. 331-334: Nature Genetics; 2019.
12. Fröhling S, Döhner H. Chromosomal Abnormalities in Cancer. *New England Journal of Medicine*. 2008;359(7):722-734.
13. Groschel S, Sanders MA, Hoogenboezem R, et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell*. 2014;157(2):369-381.
14. Yamazaki H, Suzuki M, Otsuki A, et al. A remote GATA2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating EVI1 expression. *Cancer Cell*. 2014;25(4):415-427.
15. Hsu AP, Sampaio EP, Khan J, et al. Mutations in GATA2 are associated with the autosomal dominant and sporadic monocytopenia and mycobacterial infection (MonoMAC) syndrome. *Blood*. 2011;118(10):2653-2655.
16. Hsu AP, Johnson KD, Falcone EL, et al. GATA2 haploinsufficiency caused by mutations in a conserved intronic element leads to MonoMAC syndrome. *Blood*. 2013;121(19):3830-3837.
17. Hahn CN, Chong C-E, Carmichael CL, et al. Heritable GATA2 mutations associated with familial myelodysplastic syndrome and acute myeloid leukemia. *Nature Genetics*. 2011;43:1012.
18. Ostergaard P, Simpson MA, Connell FC, et al. Mutations in GATA2 cause primary lymphedema associated with a predisposition to acute myeloid leukemia (Emberger syndrome). *Nature Genetics*. 2011;43:929.
19. Rodrigues NP, Janzen V, Forkert R, et al. Haploinsufficiency of GATA-2 perturbs adult hematopoietic stem-cell homeostasis. *Blood*. 2005;106(2):477-484.

20. International Standing Committee on Human Cytogenomic N, McGowan-Jordan J, Simons A, Schmid M. ISCN : an international system for human cytogenomic nomenclature (2016); 2016.
21. Gröschel S, Lugthart S, Richard FS, et al. High EVI1 Expression Predicts Outcome in Younger Adult Patients With Acute Myeloid Leukemia and Is Associated With Distinct Cytogenetic Abnormalities. *Journal of Clinical Oncology*. 2010;28(12):2101-2107.
22. Valk PJM, Verhaak RGW, Beijen MA, et al. Prognostically Useful Gene-Expression Profiles in Acute Myeloid Leukemia. *New England Journal of Medicine*. 2004;350(16):1617-1628.
23. Srebniak M, Boter M, Oudesluijs G, et al. Application of SNP array for rapid prenatal diagnosis: implementation, genetic counselling and diagnostic flow. *European Journal Of Human Genetics*. 2011;19:1230.
24. Srebniak MI, Diderich KEM, Joosten M, et al. Prenatal SNP array testing in 1000 fetuses with ultrasound anomalies: causative, unexpected and susceptibility CNVs. *European Journal Of Human Genetics*. 2015;24:645.
25. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.
26. Chen K, Wallis JW, McLellan MD, et al. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods*. 2009;6:677.
27. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nature Biotechnology*. 2011;29:24.
28. Mack EKM, Marquardt A, Langer D, et al. Comprehensive genetic diagnosis of acute myeloid leukemia by next-generation sequencing. *Haematologica*. 2019;104(2):277-287.
29. Gerhardt TM, Schmahl GE, Flotho C, Rath AV, Niemeyer CM. Expression of the Evi-1 gene in haemopoietic cells of children with juvenile myelomonocytic leukaemia and normal donors. *British Journal of Haematology*. 1997;99(4):882-887.
30. Privitera E, Longoni D, Brambillasca F, Biondi A. EVI-1 gene expression in myeloid clonogenic cells from juvenile myelomonocytic leukemia (JMML). *Leukemia*. 1997;11(12):2045-2048.
31. Nucifora G, Laricchia-Robbio L, Senyuk V. EVI1 and hematopoietic disorders: History and perspectives. *Gene*. 2006;368:1-11.
32. Lin P, Medeiros LJ, Yin CC, Abruzzo LV. Translocation (3;8)(q26;q24): a recurrent chromosomal abnormality in myelodysplastic syndrome and acute myeloid leukemia. *Cancer Genetics and Cytogenetics*. 2006;166(1):82-85.
33. Lennon PA, Abruzzo LV, Medeiros LJ, et al. Aberrant EVI1 expression in acute myeloid leukemias associated with the t(3;8)(q26;q24). *Cancer Genetics and Cytogenetics*. 2007;177(1):37-42.
34. De Braekeleer M, Guéganic N, Tous C, et al. Breakpoint heterogeneity in (2;3)(p15?23;q26) translocations involving EVI1 in myeloid hemopathies. *Blood Cells, Molecules, and Diseases*. 2015;54(2):160-163.
35. Trubia M, Albano F, Cavazzini F, et al. Characterization of a recurrent translocation t(2;3)(p15–22;q26) occurring in acute myeloid leukaemia. *Leukemia*. 2006;20(1):48-54.
36. Storlazzi CT, Anelli L, Albano F, et al. A novel chromosomal translocation t(3;7)(q26;q21) in myeloid leukemia resulting in overexpression of EVI1. *Annals of Hematology*. 2004;83(2):78-83.
37. Chacon D, Beck D, Perera D, Wong JWH, Pimanda JE. BloodChIP: a database of comparative genome-wide transcription factor binding profiles in human blood cells. *Nucleic Acids Research*. 2013;42(D1):D172-D177.
38. Watt SM, Chan JYH. CD164-A Novel Sialomucin on CD34+ Cells. *Leukemia & Lymphoma*. 2000;37(1-2):1-25.
39. Pellin D, Loperfido M, Baricordi C, et al. A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nature Communications*. 2019;10(1):2395.
40. Zannettino ACW, Bühring H-J, Niutta S, Watt SM, Benton MA, Simmons PJ. The Sialomucin CD164 (MGC-24v) Is an Adhesive Glycoprotein Expressed by Human Hematopoietic Progenitors and Bone Marrow Stromal Cells That Serves as a Potent Negative Regulator of Hematopoiesis. *Blood*. 1998;92(8):2613-2628.

41. Yin AH, Miraglia S, Zanjani ED, et al. AC133, a Novel Marker for Human Hematopoietic Stem and Progenitor Cells. *Blood*. 1997;90(12):5002-5012.
42. Matsuo Y1 AT, Tsubota T, Imanishi J, Minowada J. Establishment and characterization of a novel megakaryoblastic cell line, MOLM-1, from a patient with chronic myelogenous leukemia. *Human Cell*. 1991.
43. Pott S, Lieb JD. What are super-enhancers? *Nat Genet*. 2015;47(1):8-12.
44. Lovén J, Hoke Heather A, Lin Charles Y, et al. Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers. *Cell*. 2013;153(2):320-334.
45. Whyte Warren A, Orlando David A, Hnisz D, et al. Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell*. 2013;153(2):307-319.
46. Mucenski ML, Taylor BA, Ihle JN, et al. Identification of a common ecotropic viral integration site, Evi-1, in the DNA of AKXD murine myeloid tumors. *Molecular and cellular biology*. 1988;8(1):301-308.
47. Ott MG, Schmidt M, Schwarzwaelder K, et al. Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 or SETBP1. *Nature Medicine*. 2006;12(4):401-409.
48. Katayama S, Suzuki M, Yamaoka A, et al. GATA2 haploinsufficiency accelerates EVI1-driven leukemogenesis. *Blood*. 2017;130(7):908-919.

SUPPLEMENTARY INFORMATION

SUPPLEMENTARY METHODS

Differential expression analysis

Salmon[1] was used to quantify expression of individual transcripts, which were subsequently aggregated to estimate gene-level abundances with tximport[2]. Human gene annotation derived from RefSeq[3] was downloaded from UCSC[4] (RefGene) as a GTF file. Both gene- and transcript-level abundances were normalized to counts per million (CPM) for visualization in the figures of this paper. Differential gene expression analysis of count estimates from Salmon was performed with DEseq2[5]. As control, in house RNA-seq data of a cohort representative of the genetic diversity of AML cases was used (referred to as non-3q26 AML).

Allele-specific expression

To discriminate expression from different alleles, single nucleotide variants (SNVs) were first detected at the DNA level, using either whole-genome or exome sequencing data if available, or 3q-capture sequencing data otherwise. This step was performed with a custom script that integrated variants called by multiple software tools, including HaplotypeCaller and MuTecT2 from GATK[6], VarScan2[7] and bcftools[8]. The combined list of SNVs was subjected to stringent filtering to remove low-quality positions, considering the following criteria: a) strand bias, b) sequencing depth, c) alignment and base calling score, d) mappability. A highly optimized in-house tool (*annotateBamStatistics*) was then used to compute DNA and RNA allele-specific read counts at every SNV position from their respective alignment (BAM) files. For every gene, counts from all SNVs were summed to create a 2x2 contingency table (variables MAJOR/MINOR and DNA/RNA) and a χ^2 test of independence was conducted. Finally, skewed expression was determined for genes with False Discovery Rate (FDR) < 0.05 and RNA minor allele frequency < 0.35. The results were validated by visual examination of the DNA-seq and RNA-seq BAM files in IGV[9].

Copy number variant (CNV) analysis in 3q-capture data

CNV analysis was performed with CNVkit[10] in two steps. First, a pooled reference was generated based on all the 3q-capture datasets, which averaged out possible differences between them. As suggested by the instructions of the program, 5 kb regions of poor mappability were excluded from the analysis. Subsequently, the reference was employed to compute log2 copy ratios and infer discrete copy number segments using the default settings of CNVkit. Finally, we derived absolute integer copy numbers of these segments with the function “cnvkit call”. Regions with a copy number other than 2 in the vicinity of *GATA2* or *MECOM* were subjected to further scrutiny in the BAM file of the corresponding

sample: depth and variant allele frequency (VAF) were visually checked using IGV to confirm the CNV reported by CNVkit.

ChIP-seq data and analysis

H3K27ac ChIP-seq data from the inv(3) cell line MOLM-1 was previously generated in our group and is publicly available[11]. Briefly, reads were aligned to the human reference genome build hg19 with BBMap[12] and bigwig files were generated for visualization with bedtools genomecov [13] and UCSC bedGraphToBigWig[14]. Putative super-enhancers were identified on the basis of the ranked H3K27ac signal with ROSE[15], with 5kb stitching distance and excluding peaks in promoter regions. Transcription factor binding profiles (ChIP-seq) in human CD34+ cells were retrieved from the BloodChip database[16] in bigwig format. These tracks were visualized using IGV combined with Molm-1 derived H3K27ac signal to infer the presence of myeloid driven putative super-enhancers.

Bisulfite sequencing

To investigate if skewed *GATA2* expression was due to methylation of the promoter at one allele, bisulfite-sequencing experiments were performed like previously described[17]. Three regions in the *GATA2* locus were incorporated in the experiments. Based on RNA-seq data we saw that two main isoforms of *GATA2* were expressed in this patient cohort: a long and a short transcript, the latter expressed the highest. For both forms, sequences in the promoter regions were analyzed. In addition, we sequenced a region upstream of *GATA2* marked by H3K4me3 in MOLM-1 cell line. Chromosomal coordinates and primers are indicated below.

'H3K4me3 region'	**Fw:** AGCCTCTGCAGCTGGGACAAGGATGT
Chr.3: 128497666-128497881	**Rv:** GGGATTAGCTCATCTCCAGGCAGGT
'Long form *GATA2*'	**Fw:** GAGCCCCAAAGGTAGGGGCCACAGGG
Chr.3: 128492783-128492961	**Rv:** GCCTGGAGTAGAGCTGGGAGCAGG
'Short form *GATA2*'	**Fw:** GGGTAGGAGCTGGGGGTAGA
Chr.3: 128487826-128488155	**Rv:** CACCACTAAGGGACCCTCACCCCAAGG

SUPPLEMENTARY FIGURES

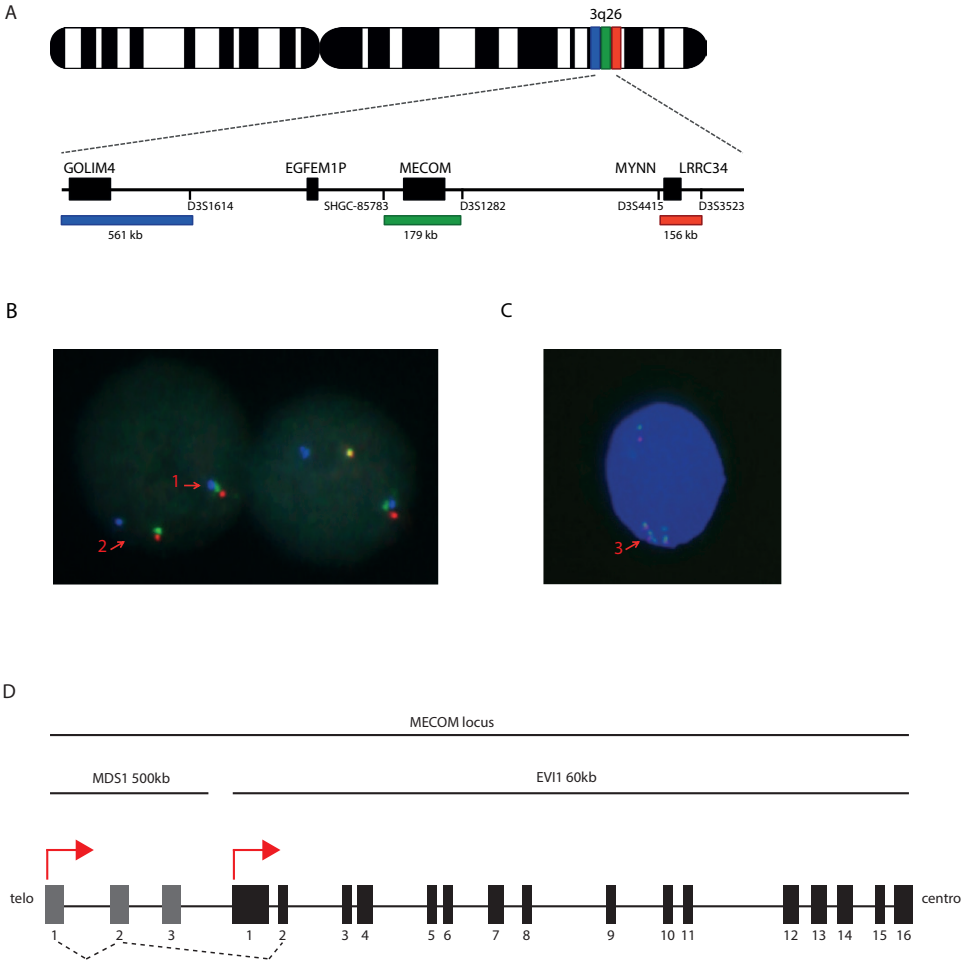


Figure S1. Schematic overview of the *MECOM* locus and *MECOM* FISH. (A) Schematic overview of *MECOM* breakapart FISH (Cytocell, figure modified from manufacturer website). (B) Example of *MECOM* FISH of AML cells with an inv(3)(q21q26). Arrow 1 indicates a normal allele with three probes. Separation of the blue probe from the green/red probes (arrows 2) recognizes the rearranged allele. (C) Example of a FISH experiment showing *MECOM* amplification (arrow 3). (D) Schematic overview of the *MECOM* locus, showing the exons of the long form *MDS1-EVI1* and the short form *EVI1* (adapted from Odera et al., 2017[18]).

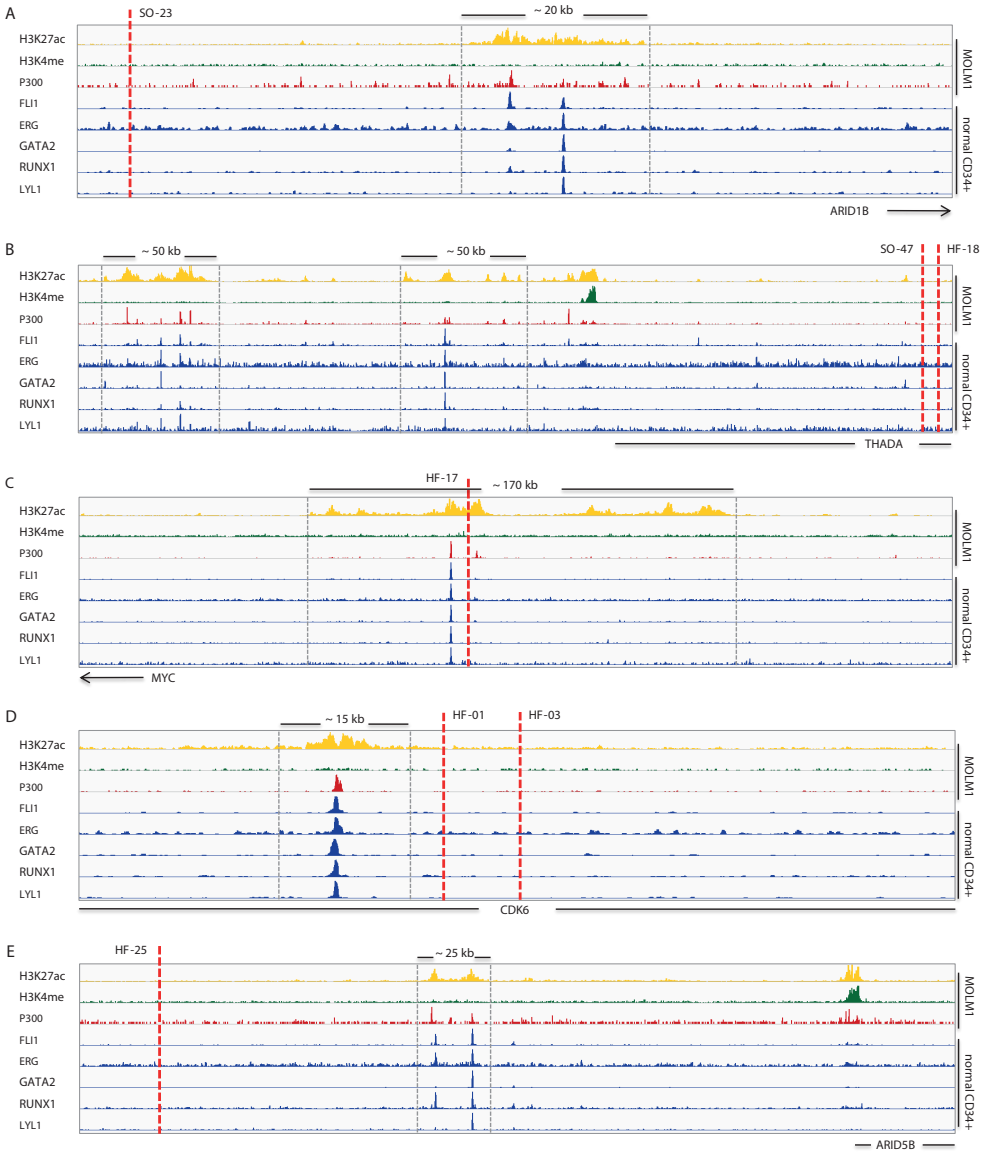


Figure S2. Super-enhancers translocated to *EVI1* in atypical 3q26. Potential super-enhancer regions (based on Figure 3G) in translocated loci, indicated by the presence of a large region of H3K27ac (yellow track), P300 (red track) and the absence of H3K4me1 (green track) (ChIP-seq data MOLM-1) and strong binding of early myeloid transcription factors, FLI1, ERG, GATA2, RUNX1 and LYL1 (blue track, ChIP-seq data CD34+ cells[16]). In each panel the translocated region that is brought in to close proximity of *EVI1* and the gene thought to be involved are depicted, the dashed red line indicates the chromosomal breakpoint. (A) #SO-23 t(3;6)(q26;q25), *ARID1B*. (B) #SO-47 and #HF-18, t(2;3)(p21;q26), *THADA*. (C) #HF-17, t(3;8)(q26;q24), *MYC*. (D) #HF-01, #HF-03, t(3;7)(q26;q11), *CDK6*. (E) #HF-25, t(3;10)(q26;q21), *ARID5B*.

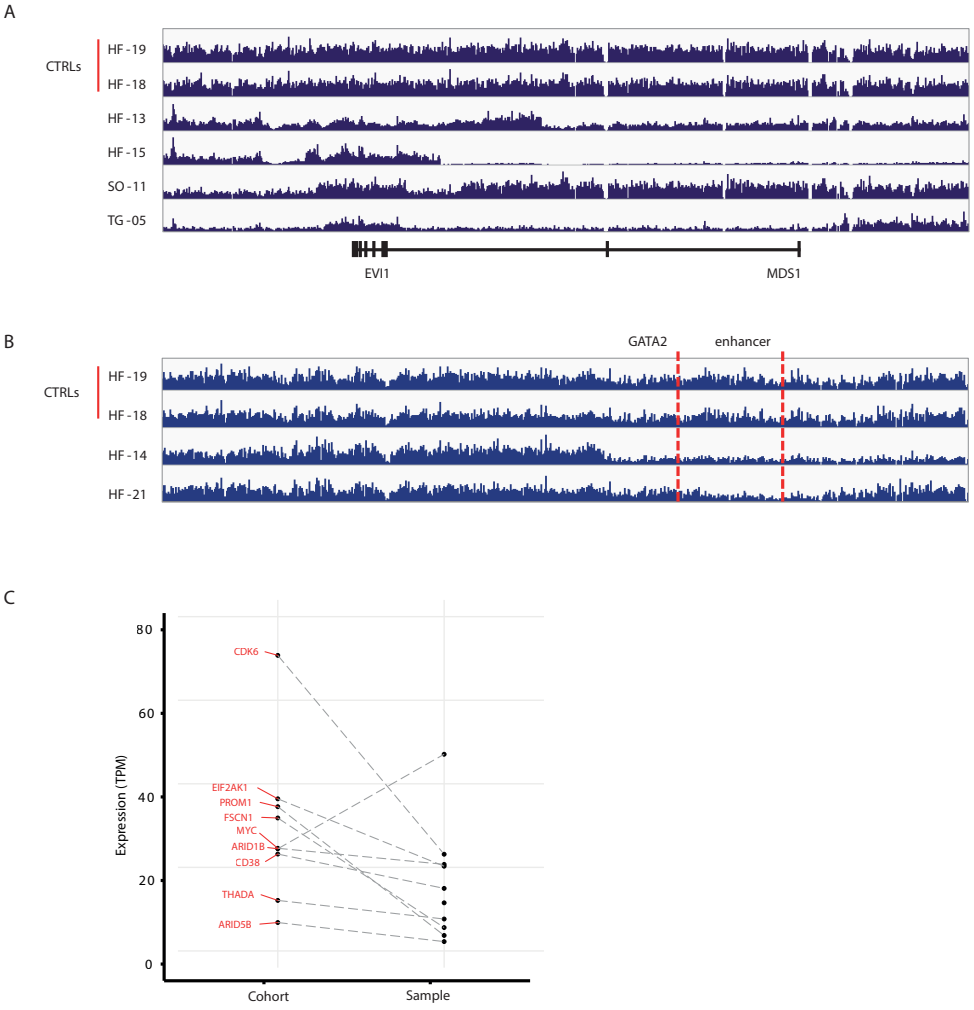


Figure S3. Copy number variants in *MECOM/GATA2* locus and decreased expression of genes in translocated region. IGV screenshots of 3q-capture tracks of atypical 3q26 cases with clear gains or losses in either the *MECOM* locus (A) or *GATA2* locus (B). The first two tracks of each panel show cases with a balanced 3q21 and 3q26 region (CTRLs). (A) In #HF-13 and #HF-15 clear gains of *EVI1* are observed but not in *MDS1*. In #SO-11 losses 3'and 5'of *EVI1* are found, but *EVI1* remains balanced. #TG-05, losses of the *MDS1* exons and the region 3' of *EVI1* are observed. Again, exactly *EVI1* remains intact. (B) In #HF-14 a loss of the *GATA2* gene and the region where the enhancer is found is lost. In patient #HF-21 the *GATA2* gene itself is balanced, but the enhancer is lost. Data of the cases shown here in S3A and S3B verifies what is observed in the SNP-array data, shown in figure 2 and 4C respectively. (C) Comparison of expression levels of the genes present in translocated regions. For each gene involved in a translocation, the first column shows its average expression in the whole cohort, whereas the second column shows its average expression in samples that carry that translocation.

REFERENCES

1. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. 2017;14(4):417-419.
2. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res*. 2015;4:1521.
3. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44(D1):D733-745.
4. Karolchik D, Hinrichs AS, Furey TS, et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. 2004;32(Database issue):D493-496.
5. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550.
6. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303.
7. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22(3):568-576.
8. Ostergaard P, Simpson MA, Connell FC, et al. Mutations in GATA2 cause primary lymphedema associated with a predisposition to acute myeloid leukemia (Emberger syndrome). *Nature Genetics*. 2011;43:929.
9. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nature Biotechnology*. 2011;29:24.
10. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol*. 2016;12(4):e1004873.
11. Groschel S, Sanders MA, Hoogenboezem R, et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell*. 2014;157(2):369-381.
12. Bushnell B. BBMap short-read aligner, and other bioinformatics tools; 2016.
13. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-842.
14. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*. 2010;26(17):2204-2207.
15. Whyte WA, Orlando DA, Hnisz D, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*. 2013;153(2):307-319.
16. Chacon D, Beck D, Perera D, Wong JWH, Pimanda JE. BloodChIP: a database of comparative genome-wide transcription factor binding profiles in human blood cells. *Nucleic Acids Research*. 2013;42(D1):D172-D177.
17. Wouters BJ, Jordà MA, Keeshan K, et al. Distinct gene expression profiles of acute myeloid/T-lymphoid leukemia with silenced CEBPA and mutations in NOTCH1. *Blood*. 2007;110(10):3706-3714.
18. Maicas M, Vázquez I, Alis R, et al. The MDS and EVI1 complex locus (MECOM) isoforms regulate their own transcription and have different roles in the transformation of hematopoietic stem and progenitor cells. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*. 2017;1860(6):721-729.

CHAPTER 4

The leukemic oncogene *EVI1* hijacks a *MYC* super-enhancer by CTCF-facilitated loops

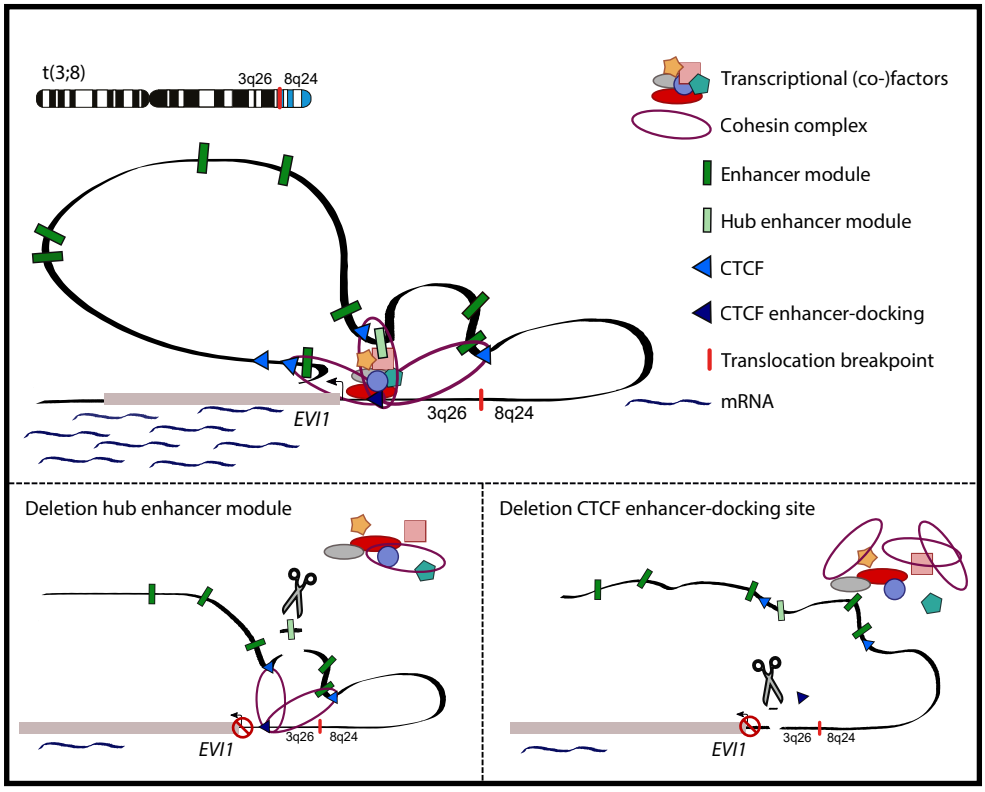
Sophie Ottema[1,2,*], **Roger Mulet-Lazaro[1,2,*]**, Claudia Erpelinck-Verschueren[1,2,*], Stanley van Herk[1,2], Marije Havermans[1,2], Andrea Arricibita Varea[1,2], Michael Vermeulen[1], H. Berna Beverloo[3], Stefan Gröschel[4,5], Torsten Haferlach[6], Claudia Haferlach[6], Bas Wouters[1,2], Eric Bindels[1], Leonie Smeenk[1,2,**], and Ruud Delwel[1,2,**]

[1] Department of Hematology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands
[2] Oncode Institute, Erasmus University Medical Center, Rotterdam, The Netherlands
[3] Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, The Netherlands
[4] German Cancer Research Center, A380, Heidelberg, Germany
[5] Department of Internal Medicine V, Heidelberg University Hospital, Heidelberg, Germany
[6] Munich Leukemia Laboratory, Munich, Germany

* These authors contributed equally: Sophie Ottema, Roger Mulet-Lazaro, Claudia Erpelinck-Verschueren
** These authors jointly supervised this work: Leonie Smeenk, Ruud Delwel

Running title: Hijacking of the MYC SE by EVI1

Nat Commun. 2021 Sep 28;12(1):5679. doi: 10.1038/s41467-021-25862-3.
PMID: 34584081


t(3;8)
3q26
8q24
Transcriptional (co-)factors
Cohesin complex
Enhancer module
Hub enhancer module
CTCF
CTCF enhancer-docking
Translocation breakpoint
mRNA
EVI1
3q26
8q24
Deletion hub enhancer module
EVI1
3q26
8q24
Deletion CTCF enhancer-docking site
EVI1
3q26
8q24

ABSTRACT

Chromosomal rearrangements are a frequent cause of oncogene deregulation in human malignancies. Overexpression of *EVI1* is found in a subgroup of acute myeloid leukemia (AML) with 3q26 chromosomal rearrangements, which is often therapy resistant. In AMLs harboring a t(3;8)(q26;q24), we observed the translocation of a MYC super-enhancer (MYC SE) to the EVI1 locus. We generated an in vitro model mimicking a patient-based t(3;8)(q26;q24) using CRISPR-Cas9 technology and demonstrated hyperactivation of EVI1 by the hijacked MYC SE. This MYC SE contains multiple enhancer modules, of which only one recruits transcription factors active in early hematopoiesis. This enhancer module is critical for *EVI1* overexpression as well as enhancer-promoter interaction. Multiple CTCF binding regions in the *MYC* SE facilitate this enhancer-promoter interaction, which also involves a CTCF binding site upstream of the *EVI1* promoter. We hypothesize that this CTCF site acts as an enhancer-docking site in t(3;8) AML. Genomic analyses of other 3q26-rearranged AML patient cells point to a common mechanism by which *EVI1* uses this docking site to hijack enhancers active in early hematopoiesis.

KEY POINTS

- A t(3;8) AML model was generated in vitro using CRISPR-Cas9
- The *MYC* super-enhancer hyperactivates *EVI1* expression in t(3;8) AML
- A single hematopoietic enhancer module is critical for *EVI1* expression
- CTCF-binding site upstream of *EVI1* hijacks enhancers in 3q26-rearranged AML

INTRODUCTION

The expression of cell lineage specific genes is highly regulated. Specific enhancer-promoter interactions and transcription factor binding to regulatory elements delineate gene expression profiles that define cell identity and function [1]. Physical interactions between enhancers and promoters primarily occur within chromosome segments enclosed by chromatin loops known as topologically associated domains (TADs) [2]. TADs are separated from each other by boundaries typically containing convergent CTCF (CCCTC-binding factor) occupied sites [3]. According to the loop extrusion model, the cohesin complex catalyzes the formation of loops and CTCF dimers act as anchors to these loops [4]. CTCF and the cohesin complex, but also other factors like Ying Yang 1 (YY1), may also contribute to enhancer-promoter looping [5-8]. However, not all promoters or enhancers within a TAD interact with each other. The mechanisms by which promoters interact with certain enhancers and not with others are not fully understood [9,10]. Transcriptional control of genes driven by particular enhancer-promoter combinations depends on the availability of transcription factors and their ability to bind specific regulatory elements [8,11].

Chromosomal rearrangements frequently lead to changes in the expression or function of genes causing malignant transformation [12]. Often breakpoints are found within gene bodies, resulting in fusion oncogenes driving tumorigenesis [13]. Alternatively, when a regulatory element of a certain gene is translocated into the vicinity of another gene, it can lead to deregulation of both the donor and the acceptor genes. Well-described examples are the inv(3)(q21q26) or t(3;3)(q21;q26) rearrangements in acute myeloid leukemia (inv(3)/t(3;3) AML), in which a *GATA2* enhancer at 3q21 is hijacked by *EVI1* at 3q26, causing *EVI1* overexpression and *GATA2* haploinsufficiency [14,15]. AML is a heterogeneous disease, with *EVI1* positive (*EVI1*+) inv(3)/t(3;3) patients being identified as a subgroup with a very poor response to therapy [16-19]. Besides inv(3)/t(3;3), many other *EVI1*+ AML cases with 3q26 rearrangements have been reported, including translocations t(2;3)(p21;q26), t(3;7)(q26;q24), t(3;6)(q26;q11) and t(3;8)(q26;q24) [18,20-27]. We hypothesize that in all these rearrangements *EVI1* overexpression is induced by the repositioning of an enhancer that can interact with the *EVI1* promoter, as shown for inv(3)/t(3;3) AML [14,15]. We performed targeted next generation sequencing (NGS) of the long arm of chromosome 3 (3q-seq) in translocation t(3;8)(q26;q24) AML harboring an *EVI1/MYC* rearrangement [22,27]. Applying CRISPR-Cas9 technology, we generated a human t(3;8) cell line model with an *eGFP* reporter cloned 3' of *EVI1*. This unique model was used to investigate how enhancer-promoter interactions drive oncogenic *EVI1* expression in leukemia. We demonstrate that CTCF in combination with transcription factors active in early hematopoiesis is essential in enhancer hijacking and oncogene activation.

RESULTS

MYC super-enhancer translocation and *EVI1* overexpression in t(3;8)(q26;q24) AML

Using 3q-seq, the exact chromosomal breakpoints were determined in 10 AML samples with a translocation t(3;8)(q26;q24), hereafter referred to as t(3;8) AML. All breakpoints at 3q26.2 occurred upstream of the *EVI1* promoter (Figure 1A). At chromosome 8, the breakpoints were downstream of the oncogene *MYC* at 8q24, leaving the gene intact at its original location. In all 10 cases a genomic region reported as a *MYC* super-enhancer (SE) had been translocated to *EVI1* (Figure 1B). The *MYC* SE harbors approximately 150 Kb of open chromatin enriched with histone mark H3K27 acetylation (H3K27ac) and is located 1.7 Mb downstream of *MYC* (Figure 1B). This locus has been reported to be essential for transcriptional control of *MYC* expression in normal hematopoiesis [28]. H3K27ac determined by ChIP-seq revealed *EVI1* promoter activity in t(3;8) AML patient cells, comparable to the promoter activity in AML with inv(3)(q21q26). H3K27ac was absent at the *EVI1* promoter in *EVI1* negative (*EVI1*$^{-}$) non-3q26 AML (Figure 1A, lower panel). Accordingly, *EVI1* expression was found to be highly elevated in t(3;8) compared to non-3q26 rearranged AMLs (Figure 1C). The *EVI1* levels in t(3;8) AMLs were comparable to the levels found in AMLs with inv(3)/t(3;3). These data support the hypothesis that *EVI1* overexpression in t(3;8) AML is caused by the translocation of the *MYC* SE.

A t(3;8) cell model recapitulates *EVI1* overexpression in human AML

To study the transcriptional activation of *EVI1* by the *MYC* SE, we generated a human myeloid cell model with a translocation t(3;8)(q26;q24). We introduced e*GFP* in frame with a T2A self-cleavage site downstream of *EVI1* in K562 cells (Figure 2A). Successful integration of the insert is shown for two clones by flow cytometry and PCR (Figure 2B, Supplementary Fig. 1A-C). Decreased eGFP levels were observed in the K562 *EVI1-eGFP* model after shRNA directed *EVI1* knockdown (Figure 2C-D and Supplementary Fig. 1D-G). Next, sgRNAs for CRISPR-Cas9 editing were designed based on the genomic breakpoints of one of the t(3;8) AML patients in our cohort (Figure 1A). Double strand DNA breaks were generated at 3q26 and 8q24 (Figure 2E) using those guides. We hypothesized that the translocated *MYC* SE can activate *EVI1* transcription, which consequently leads to increased *eGFP* levels. As shown in Figure 2F, less than 0.1% of the sgRNA-treated K562 *EVI1-eGFP* cells showed increased eGFP levels. After two consecutive rounds of FACS sorting in combination with cell culture expansion, we obtained 95% eGFP positive cells of which single clones were isolated by single cell sorting (process done similarly for both clones 8 and 24, Figure 2F shows clone 24). The presence of a t(3;8) was demonstrated for four of these clones by PCR (Clone 24-7, Figure 2H) and Sanger sequencing (Supplementary Fig. 2A). A combination of three separate diagnostic FISH probes for *MECOM*, *MYC* and centromere chromosome 8 confirmed the

successful generation of a translocation t(3;8) in all four clones (Supplementary Fig. 2B-E). The translocation caused a strong increase of mRNA and protein levels of *EVI1* as well as of *eGFP* expression (Figure 2G, J, K). No significant difference in *MYC* expression was observed between the parental K562 *EVI1-eGFP* and t(3;8) clones (Figure 2I). Upon *EVI1* knockdown by shRNA, *eGFP* and *EVI1* expression were reduced as shown for clone 24-7 and 8-4 (Figure 2L-M and Supplementary Fig. 2F-G). We conclude that eGFP is a sensitive and reliable marker for *EVI1* expression in this *EVI1-eGFP* t(3;8) model, and that the translocated *MYC* SE strongly enhances *EVI1* transcription.

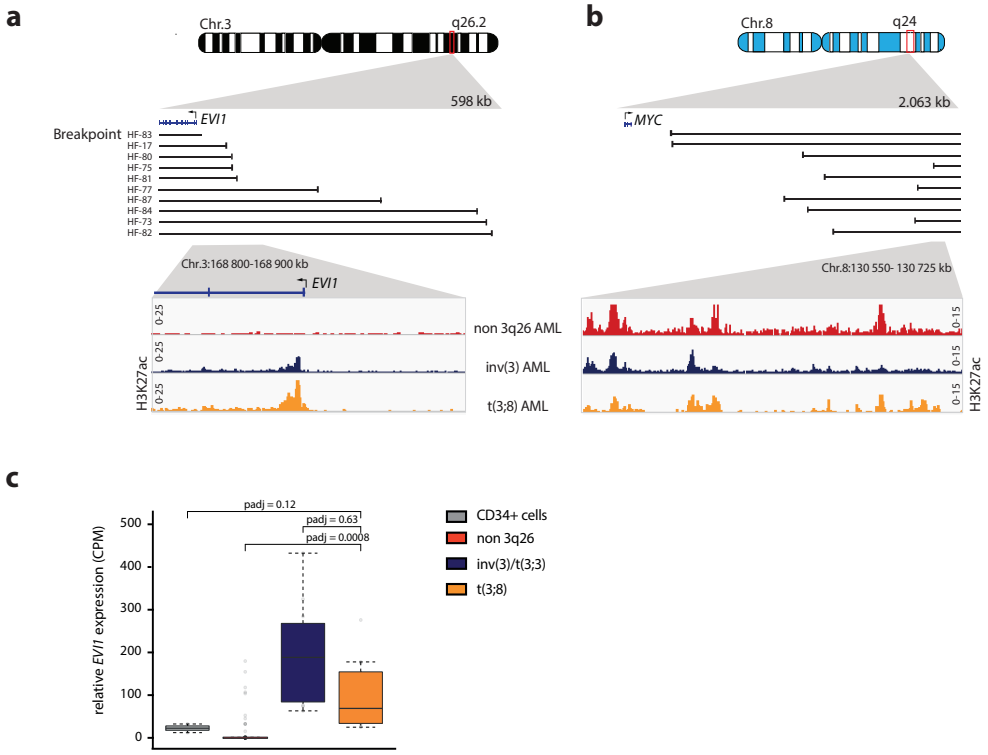


Figure 1. *MYC* super-enhancer translocation and *EVI1* overexpression in t(3;8)(q26;q24) AML. (a) Upper part, schematic depiction of Chr.3, zoomed in on 3q26.2. Black lines correspond to sample specific breakpoints detected by 3q-seq for each indicated t(3;8)(q26;q24) patient. Lower part: zoom-in on the *EVI1* promoter, H3K27ac ChIP-seq data for a primary non-3q26 AML sample in red (N=1, AML-185), an inv(3)(q21q26) in blue (N=1, AML-2190) and a t(3;8)(q26;q24) in orange (N=1, AML-17). (b) Similar to A, but here in the upper part a schematic depiction of Chr.8, zoomed in on 8q24. Lower part: H3K27ac ChIP-seq data as in A, but here a zoom-in on the +1.7 Mb *MYC* super-enhancer. (c) *EVI1* expression measured by RNA-seq in counts per million (CPM) for normal CD34+ HSPCs (N=9, grey), non-3q26 AMLs (N=114, red), inv(3)/t(3;3)(q21;q26) AMLs (N=11, blue), and t(3;8)(q26;q24) AMLs (N=10, orange). The lower and upper edges of the boxplots represent the first and third quartiles, respectively, the horizontal line inside the box indicates the median. The whiskers extend to the most extreme values within the range comprised between the median and 1.5 times the interquartile range. The circles represent outliers outside this range. Statistical significance of the comparisons between these groups was determined by the Wald test in the DESeq2 package. Adjusted p-values (padj) following multiple testing correction by the Benjamini–Hochberg procedure are displayed.

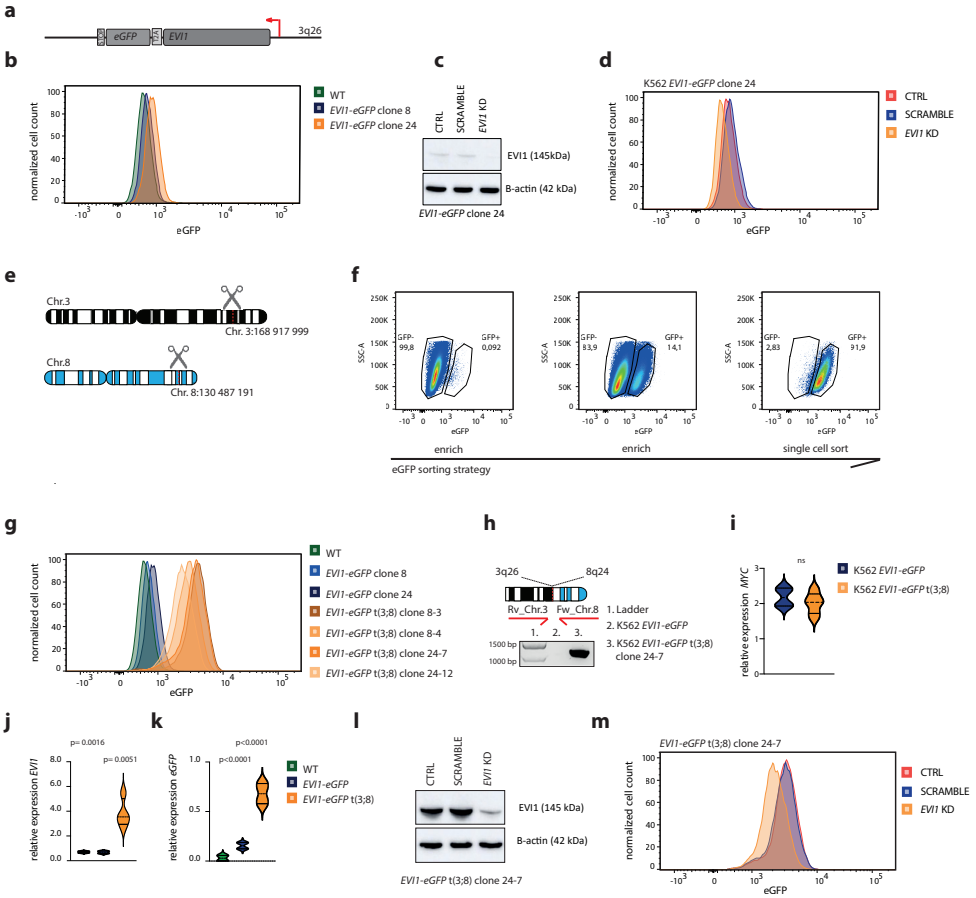


Figure 2. A t(3;8) cell model recapitulates *EVI1* overexpression in human AML. (a) Schematic overview of *EVI1-T2A-eGFP*. (b) Flow cytometry plot presenting eGFP levels in K562-*EVI1-eGFP* clones. (c) Western blot show EVI1 levels after shRNA directed *EVI1* knockdown (KD), compared to the control and scrambled shRNA in K562 *EVI1-eGFP* clone 24. Source data are provided as a Source Data file. (d) Flow cytometry plot presenting eGFP after *EVI1* knockdown (KD), compared to the control and scrambled shRNA in K562 *EVI1-eGFP* clone 24. (e) Schematic overview of the generation of a t(3;8)(q26;q24) in vitro using CRISPR-Cas9 technology, referred to in short as t(3;8). (f) Sorting strategy to enrich twice for cells with high *EVI1-eGFP* expression and select eGFP positive single clones with a t(3;8). (g) Flow cytometry plot presenting eGFP levels in t(3;8) K562 clones compared to the parental K562 *EVI1*-eGFP clones. Two parental clones (8 and 24), and four t(3;8) clones (8-3, 8-4, 24-7 and 24-12) are shown. (h) PCR amplicon covering the 3q26;8q24 breakpoint K562 *EVI1-eGFP* cells harboring a t(3;8), PCR for all single t(3;8) clones are provided in the Source Data file. (i) No significant difference in *MYC* expression (relative to *PBGD* expression) was observed between the K562 *EVI1-eGFP* parental clones (8 and 24) and the K562 *EVI1-eGFP* t(3;8) clones (8-3, 8-4, 24-7 and 24-12). Statistical test: ordinary one-way ANOVA (ns = not significant). The error bar represents the standard deviation (SD). (j) Significant higher *EVI1* expression (relative to *PBGD* expression) shown by qPCR in the t(3;8) (N=4) clones, compared to the parental clones (N=2, P=0.0016) and WT K562 (P=0.0051). Statistical test: ordinary one-way ANOVA. The error bar represents the standard deviation (SD). (k) *eGFP* expression relative to *PBGD* shown by qPCR in the t(3;8) (N=4) clones, compared to the parental clones (N=2, P<0.0001) and WT K562 (P<0.0001). Statistical test: ordinary one-way ANOVA. The error bar represents the standard deviation (SD). (l) Western blot shows lower EVI1 levels for *EVI1* shRNA directed knockdown (KD), as compared to the control and scrambled shRNA in K562 *EVI1-eGFP* t(3;8) clone 24-7. Source data are provided as a Source Data file. (m) Flow cytometry plot presenting eGFP after *EVI1* shRNA directed knockdown (KD), as compared to the control and scrambled shRNA in K562 *EVI1-eGFP* t(3;8) clone 24-7.

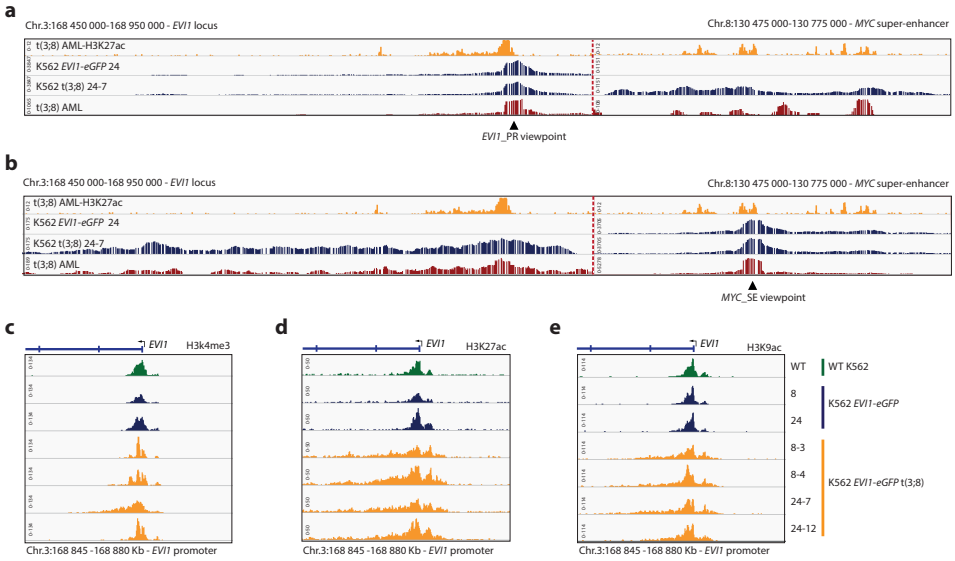


Figure 3. *EVI1* promoter hyperactivation upon interaction with *MYC* SE in t(3;8) AML. (a) Chromatin interaction shown by 4C-seq data, using the *EVI1* promoter as viewpoint (triangle symbol). The upper panel shows H3K27ac ChIP-seq data of a t(3;8) primary AML (AML-17). Indicated by H3K27ac signal peaks, on the left the *EVI1* promoter and on the right the -1.7 Mb *MYC* super enhancer, separated by a dotted red line. In the first 4C track (blue), parental K562 *EVI1-eGFP* clone 24; in the second, K562 t(3;8) *EVI1-eGFP* clone 24-7 (also blue); and in the bottom track (red), data of a primary t(3;8) AML (AML-17). (b) Similar to A, but using the *MYC* super enhancer as viewpoint (triangle symbol). The long stretch (500 kb) of chromosomal interaction shown for the K562 t(3;8) *EVI1-eGFP* clone 24-7 shows high resemblance with the interaction seen for the primary t(3;8) AML (AML-17) (second blue and red tracks respectively). (c) H3K4me3 ChIP-seq data for K562 *EVI1-eGFP* parental lines (blue), the t(3;8) clones 8-3, 8-4, 24-7, 24-12 (orange) and K562 WT (green). A peak located on the *EVI1* transcriptional start site marks the promoter region. (d) H3K27ac ChIP-seq data, comparing *EVI1* promoter activation of the four t(3;8) clones (orange) to the *EVI1-eGFP* parental lines (blue) or K562 WT (green). (e) H3K9ac ChIP-seq data, confirming the hyperactivation of the *EVI1* promoter in the t(3;8) clones (orange), compared to the parental lines (blue) or K562 WT (green).

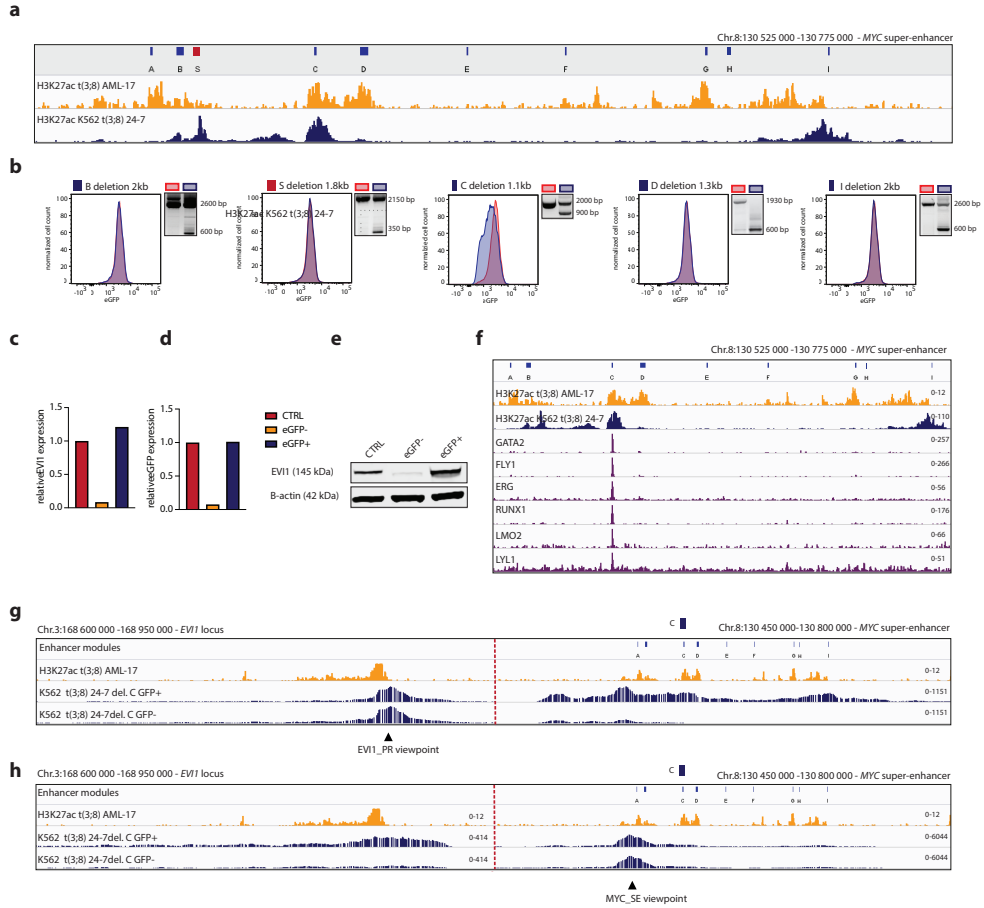


Figure 4. One critical enhancer module in the *MYC* SE drives *EVI1* transcription. (a) Overview of the *MYC* super-enhancer, with previously characterized individual enhancer modules A-I [28] and added module S based on high H3K27ac signal at this location in all K562 *EVI1-eGFP* t(3;8) clones. Underneath, H3K27ac of a primary t(3;8) AML (AML-17, orange) and of K562 t(3;8) *EVI1-eGFP* clone 24-7 (blue). (b) Flow cytometry plots (clone 24-7) shown for each indicated enhancer module deletion. In red the control cells (no Cas9), and in blue the cells carrying the deletion. On the right of each graph the successful deletion of each element is shown by PCR. Source data are provided as a Source Data file. (c) *EVI1* expression relative to *PBGD* by qPCR in eGFP- and eGFP+ sorted cell fractions after deletion of enhancer module C. The bars represent only one data point, from the exact experiment as the flow cytometry data is shown in panel B. (d) *eGFP* expression relative to *PBGD* by qPCR in eGFP- and eGFP+ sorted cell fractions after deletion of enhancer module C. The bars represent only one data point, from the exact experiment as the flow cytometry data is shown in panel B. (e) EVI1 protein levels by Western blotting in eGFP- and eGFP+ sorted fractions after deletion of enhancer module C. Source data are provided as a Source Data file. (f) *MYC* SE element C recruits a set of HSPC-active transcription factors shown by ChIP-seq data of CD34+ cells (purple tracks [29]), H3K27ac of primary t(3;8) AML (AML-17, orange) and of a K562 t(3;8) (clone 24-7, dark blue) to illustrate enhancer modules. (g) Chromatin interaction shown by 4C-seq data, using the *EVI1* promoter as viewpoint (triangle symbol). The *EVI1* promoter and the -1.7 Mb *MYC* SE are shown on the left and right sections respectively, separated by a dotted red line. The upper orange panel shows H3K27ac ChIP-seq of a t(3;8) primary AML (AML-17). In blue, 4C-seq tracks of K562 *EVI1-eGFP* t(3;8) clone 24-7 cells in which the enhancer module C was deleted. In the upper blue track, eGFP+ sorted cells, and in the lower blue track, eGFP- cells. (h) Same as (g), but using the *MYC* super-enhancer as a viewpoint (triangle symbol).

EVI1 promoter hyperactivation upon interaction with *MYC* SE in t(3;8) AML

4C-seq experiments taking the *EVI1* promoter (EVI1_PR) as a viewpoint revealed specific interaction with the *MYC* SE in *EVI1-eGFP* t(3;8) cells, which was not found in the parental K562 *EVI1-eGFP* line (clone 24-7 and clone 24 respectively, Figure 3A). This t(3;8)-specific interaction between the *EVI1* promoter and *MYC* SE was confirmed in t(3;8) clone 8-4 (Supplementary Fig. 3D) and by reciprocal 4C-seq using the *MYC* SE as a viewpoint (clone 24-7, Figure 3B). A comparable *EVI1* promoter – *MYC* SE interaction was found in a primary t(3;8) AML sample (Figure 3A-B), confirming that the K562 *EVI1-eGFP* t(3;8) model recapitulates primary AML. ChIP-seq for H3K4 trimethylation (H3K4me3, Figure 3C) indicated the presence of an active *EVI1* promoter in all K562 clones. However, H3K27 and H3K9 acetylation (H3K27ac and H3K9ac) levels were strongly increased at the promoter in all four t(3;8) clones, revealing a hyperactivated *EVI1* promoter (Figure 3D-E) upon interaction with the translocated *MYC* SE.

One critical enhancer module in the *MYC* SE drives *EVI1* transcription

The *MYC* SE is a cluster of multiple individual enhancer modules that may recruit different sets of transcription factors [28]. To investigate which of the enhancer modules are driving oncogenic *EVI1* transcription in t(3;8) AML, we designed sgRNAs to sequentially delete those individual modules. H3K27ac ChIP-seq data of a primary t(3;8) AML and of t(3;8) clone 24-7 were used to illustrate the different enhancer modules A-I described previously [28] (Figure 4A). The deletion of these modules by CRISPR-Cas9 using specific sgRNA pairs was shown by PCR and the effect on *EVI1* expression was determined by flow cytometry (Figure 4B). Only the deletion of module C caused loss of *EVI1/eGFP* expression. Due to existence of multiple alleles (K562 has trisomy 8) and partial efficiency of CRISPR-Cas9 in creating deletions, the translocated allele is exclusively targeted in a subpopulation of cells. As a consequence, not all cells lose *EVI1* expression and show a GFP shift in the flow cytometry plot. A loss of *EVI1* mRNA and EVI1 protein levels was observed in the eGFP negative sorted cell fraction when module C was deleted (Figure 4C-E and Supplementary Fig. 3A). In a control clone in which *EVI1-eGFP* expression was increased due to the amplification of *EVI1* instead of the translocation of the *MYC* SE (Supplementary Fig. 4A-E), the expression of *EVI1*-eGFP was not affected by mutating the *MYC* SE (Supplementary Fig. 4F). ATAC-seq and H3K27ac ChIP-seq in t(3;8) AML patients showed that module C was distinctly accessible and active compared to other modules (Supplementary Fig. 3B-C). Furthermore, ChIP-seq data revealed binding of early hematopoietic regulators (GATA2, FLI1, ERG, RUNX1, LMO2 and LYL1) to module C in CD34+ hematopoietic stem and progenitor cells (HSPCs) [29] (Figure 4F). Similar transcription factor binding patterns were found in t(3;8) AML patients and K562 cells, further confirming the functional significance of this module in this context (Supplementary Fig. 3C).

4C-seq taking the *EVI1* promoter as a viewpoint revealed that the strong interaction with the *MYC* SE was severely diminished in the eGFP negative fraction upon deletion of module C (Figure 4G). This loss of chromosomal interaction was also observed taking the *MYC* SE

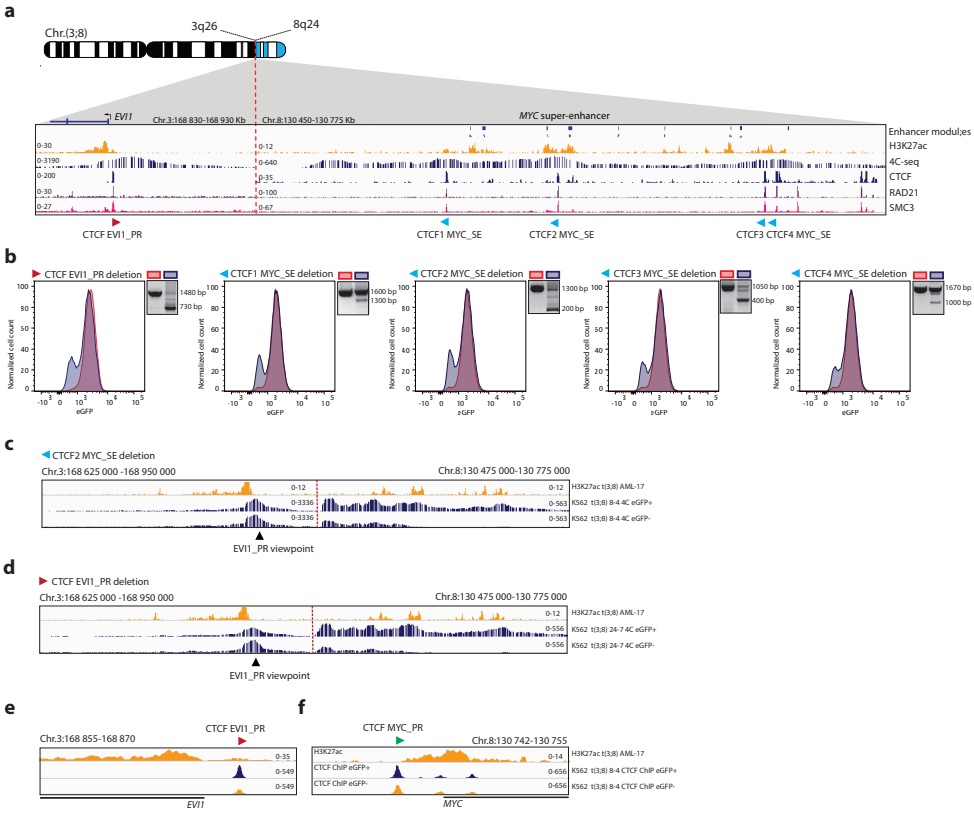


Figure 5. CTCF binding sites in *MYC* SE are involved in the interaction with the *EVI1* promoter. (a) Schematic overview of the t(3;8) model, zoom-in on the breakpoint with 3q26 on the left and 8q24 on the right (separated by a dotted red line). The *EVI1* promoter and the *MYC* super-enhancer are illustrated by ChIP-seq H3K27ac data of primary t(3;8) AML (AML-17) in orange on top. Below in blue, 4C-seq data showing the interaction pattern between the *EVI1* promoter and the *MYC* enhancer modules in K562 t(3;8) clone 24-7. The lower 3 tracks show ChIP-seq data of CTCF (blue in K562 t(3;8) clone 24-7) and the cohesin subunits RAD21 (purple) and SMC3 (pink), both in K562 WT and retrieved from ENCODE [63]. (b) Flow cytometry plot of K562 *EVI1-eGFP* t(3;8) clone 8-4 cells after deletion of the indicated CTCF binding site (blue graph), and in red the control cells (no Cas9). On the right of the graph the deletion is shown by PCR. Source data are provided as a Source Data file. (c) Chromatin interaction at the *MYC* SE in eGFP+ (upper blue track) and eGFP- cells (lower blue track), shown by 4C-seq with the *EVI1* promoter as viewpoint, after deletion of the *MYC* SE CTCF2 binding site. The top H3K27ac ChIP-seq track (orange) shows the presence of the active *EVI1* promoter and the modules of the *MYC* SE. (d) Chromatin interaction at the *MYC* SE in eGFP+ (upper blue track) and eGFP- cells (lower blue track), shown by 4C-seq with the *EVI1* promoter as viewpoint after deletion of the *EVI1* PR CTCF binding site (indicated by red arrow, corresponding to the CTCF *EVI1*_PR locus in A). The top H3K27ac ChIP-seq track (orange) shows the presence of the active *EVI1* promoter and the modules of the *MYC* SE. (e) CTCF ChIP-seq presenting CTCF occupancy in eGFP- cells (clone 8-4, orange) compared to eGFP+ cells (clone 8-4, blue) at the CTCF binding site upstream of the *EVI1* promoter after deletion of this CTCF *EVI1*_PR site. The top H3K27ac ChIP-seq track (orange) shows the presence of the active *EVI1* promoter and the modules of the *MYC* SE. (f) The same CTCF ChIP-seq tracks are shown (clone 8-4), but now presenting unchanged CTCF occupancy at the CTCF binding site upstream of the *MYC* promoter.

as a viewpoint (Figure 4H). Deletions of enhancer module D and I affected neither *EVI1* expression nor enhancer-promoter looping (Figure 4B and Supplementary Fig. 3D-E). Our data demonstrate that aberrant *EVI1* expression in t(3;8) AML depends on a single enhancer

module within the *MYC* SE that recruits a cluster of key hematopoietic transcription factors and facilitates promoter-enhancer looping.

CTCF binding sites in *MYC* SE are involved in the interaction with the *EVI1* promoter

The *EVI1* promoter interacts with the *MYC* SE over a long stretch of chromatin (275 Kb) with multiple zones of strong interaction indicative of a highly organized enhancer-promoter interaction (Figure 5A). These high interaction zones in the *MYC* SE were associated with enhancer modules, but also with CTCF/Cohesin binding based on ChIP-seq data (Figure 5A). Notably, CTCF binding motifs in the *MYC* SE are arranged in a convergent orientation to that of a CTCF binding site upstream of the *EVI1* promoter, suggesting the existence of a CTCF-facilitated enhancer-promoter loop. Using CRISPR-Cas9 technology, we sequentially deleted every CTCF binding site in the *MYC* SE. The deletions and their effect on *EVI1* expression were shown by PCR and *eGFP* flow cytometry (Figure 5B). A fraction of cells lost *eGFP* expression upon deletion of each of the CTCF binding sites in the *MYC* SE. The CTCF site closest to module C (CTCF2) was deleted and cells were sorted based on eGFP expression. A severe loss of promoter-enhancer interaction was observed in the eGFP negative cells (Figure 5C and Supplementary Fig. 5A). This strongly supports a role for CTCF/cohesin in the promoter-enhancer complex formation and maintenance, and consequently in *EVI1* regulation in t(3;8) AML.

CTCF binding site upstream of the *EVI1* promoter hijacks the *MYC* SE in t(3;8) AML

Upstream of the *EVI1* promoter a CTCF binding site in the forward orientation (CTCF EVI1_PR) was found by ChIP-seq and motif analysis (Figure 5A and 6A). Deletion of this CTCF binding region caused loss of *EVI1* expression as determined by eGFP flow cytometry. This loss of eGFP expression was comparable to the loss of expression upon deletion of the *MYC* SE CTCF sites (Figure 5B). Deletion of this CTCF site also caused a severe loss of promoter-enhancer looping in eGFP negative cells, as measured by 4C-seq (Figure 5D and Supplementary Fig. 5B). ChIP-seq showed that CTCF occupancy upstream of the *EVI1* promoter was indeed reduced upon deletion of this site (Figure 5E). CTCF occupancy at other CTCF binding sites, e.g. upstream of the *MYC* promoter (Figure 5F), was not affected. Aiming to specifically target the CTCF binding and not other transcription factor binding motifs within this genomic region, more subtle mutations were made close to the CTCF binding motif using a single sgRNA (Figure 6A). The mutations introduced by this single sgRNA strongly downregulated *eGFP/EVI1* expression (Figure 6B). A high mutation frequency was obtained in the eGFP negative sorted cells near the CTCF motif (Figure 6C-D). These mutations led to a decrease of CTCF binding specifically at this site (Figure 6E-F) and a severe loss of enhancer-promoter interaction (Figure 6G) in the eGFP negative sorted cells. Taken together, these data demonstrate an important role for the CTCF binding site upstream of the *EVI1* promoter in the hijacking of the *MYC* SE and the hyperactivation of *EVI1*.

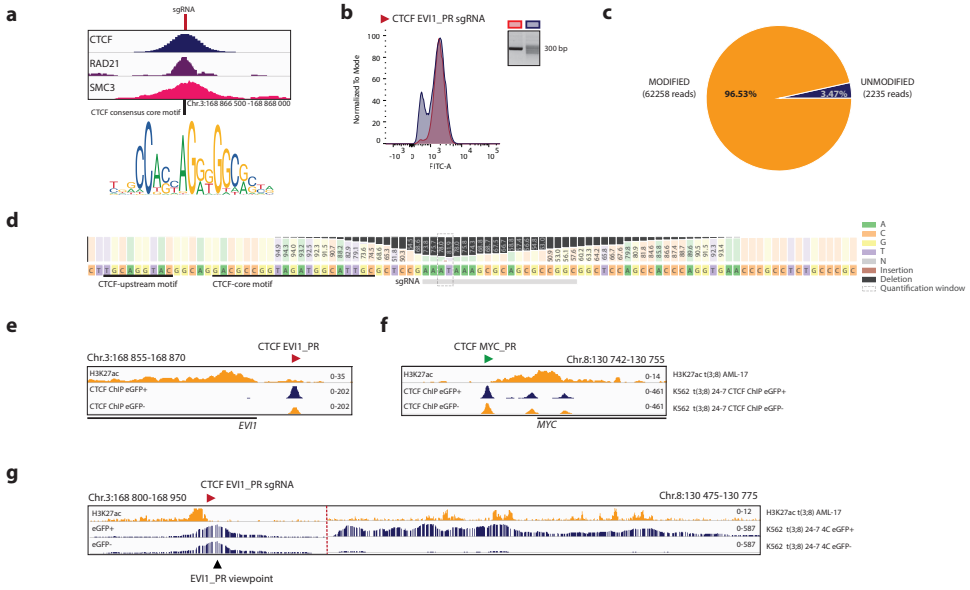
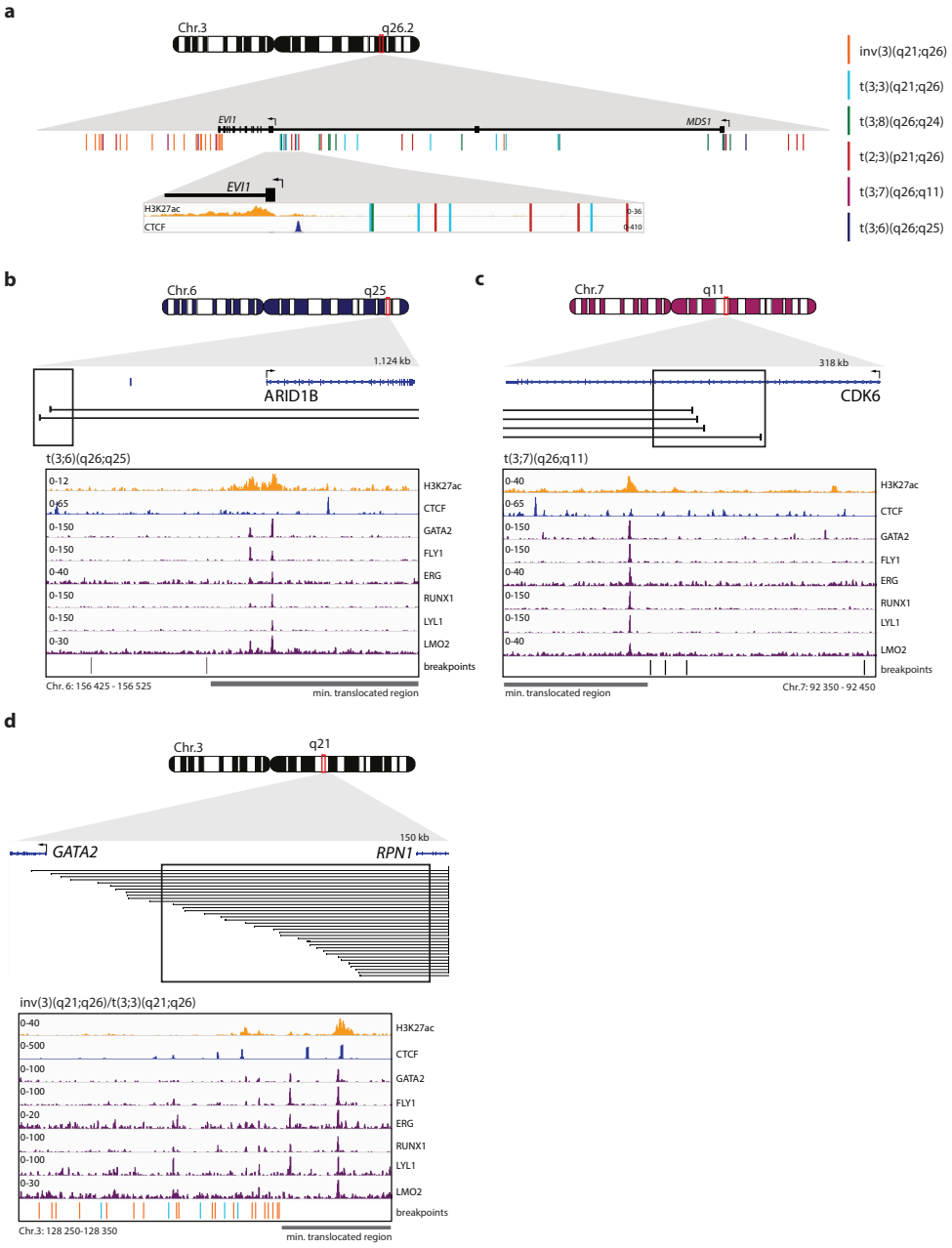


Figure 6. CTCF binding site upstream of the *EVI1* promoter hijacks a *MYC* SE in t(3;8) AML. (a) ChIP-seq data of CTCF (blue) and the cohesin subunits RAD21 (purple) and SMC3 (pink) in K562, with a zoom-in on the *EVI1* promoter binding site. The vertical line indicates the exact cleavage site of the sgRNA and the CTCF motif as described by JASPAR [64] below. (b) Flow cytometry overlay plot after targeting the CTCF *EVI1*_PR binding site by sgRNA (clone 8-4, blue graph) and in red the control cells (clone 8-4, no Cas9). On the right of the graph, the mutations introduced by the single sgRNA in the amplicon over the cutting site are shown by PCR. Source data are provided as a Source Data file. (c) Amplicon-seq data showing the percentage of modified (orange) and unmodified (blue) reads in the eGFP- sorted cell fraction after targeting CTCF *EVI1*_PR. (d) Amplicon-seq data showing the mutations in the nucleotides around the Cas9 cleavage site, in the eGFP- sorted cell fraction after targeting CTCF *EVI1*_PR with sgRNA. The bars and numbers indicate percentage of reads found with the particular mutation, below the locations of the sgRNA (grey bar) and the CTCF motifs (black lines). (e) CTCF ChIP-seq presenting CTCF occupancy in the eGFP+ (clone 24-7, blue), and eGFP- (clone 24-7, orange) fractions after targeting CTCF *EVI1*_PR with the sgRNA. The top H3K27ac track (orange) indicates the presence of an active promoter. (f) The same CTCF ChIP-seq tracks as in (e) are shown, but here presenting unchanged CTCF occupancy at the CTCF binding site upstream of the *MYC* promoter. (g) Chromatin interaction at the *MYC* SE for eGFP+ and eGFP- cells (clone 24-7), shown by 4C-seq with the *EVI1* promoter as viewpoint after targeting the CTCF motif with the sgRNA.


a
Chr.3
q26.2
EVI1
MDS1
EVI1
H3K27ac
CTCF
inv(3)(q21;q26)
t(3;3)(q21;q26)
t(3;8)(q26;q24)
t(2;3)(p21;q26)
t(3;7)(q26;q11)
t(3;6)(q26;q25)
b
Chr.6
q25
1.124 kb
ARID1B
t(3;6)(q26;q25)
0-12
0-65
0-150
0-150
0-40
0-150
0-150
0-30
H3K27ac
CTCF
GATA2
FLY1
ERG
RUNX1
LYL1
LMO2
breakpoints
Chr. 6: 156 425 - 156 525
min. translocated region
c
Chr.7
q11
318 kb
CDK6
t(3;7)(q26;q11)
0-40
0-65
0-150
0-150
0-40
0-150
0-150
0-40
H3K27ac
CTCF
GATA2
FLY1
ERG
RUNX1
LYL1
LMO2
breakpoints
min. translocated region
Chr.7: 92 350 - 92 450
d
Chr.3
q21
150 kb
GATA2
RPN1
inv(3)(q21;q26)/t(3;3)(q21;q26)
0-40
0-500
0-100
0-100
0-20
0-100
0-100
0-30
H3K27ac
CTCF
GATA2
FLY1
ERG
RUNX1
LYL1
LMO2
breakpoints
Chr.3: 128 250-128 350
min. translocated region

CTCF enhancer-docking site upstream of the *EVI1* promoter is preserved in all 3q26-rearranged AMLs

The essential role of the CTCF binding site upstream of the *EVI1* promoter in mediating the interaction with a hijacked enhancer would predict that this site remains unaffected in 3q26-rearranged AMLs. Indeed, all breakpoints of t(3;8) AMLs analyzed were found upstream of this CTCF site, placing the *MYC* SE 5' of *EVI1* (Figure 1A and 7A). In t(3;3)(q21;q26) AML the *GATA2* enhancer similarly translocates 5' of the *EVI1* promoter and of the CTCF binding site (Figure 7A) [15]. In AML with inv(3)(q21q26) the *GATA2* enhancer translocates 3' of *EVI1* [15] (Figure 7A), leaving the enhancer-interacting CTCF site in position with respect to *EVI1* as well. We collected samples from AML patients with translocations t(2;3)(p21;q26), t(3;7)(q26;q24) or t(3;6)(q26;q11) and carried out 3q-seq (Figure 7A and Supplementary Fig. 6A). Irrespective of whether a translocation had occurred 3' or 5' of *EVI1*, the CTCF binding site flanking the *EVI1* promoter was never disrupted, suggesting a key role for this binding site in this AML subtype. Accordingly, ChIP-seq revealed constitutive binding of CTCF to this location across various leukemias, including not only 3q26-rearranged AMLs, but also other AMLs and acute lymphoid leukemia (Supplementary Fig. 6B).

Enhancers of the genes *GATA2* [15] and *MYC* are respectively responsible for *EVI1* activation in inv(3)/t(3;3) and t(3;8) AML. Using 3q-seq, we observed that regions near the genes *CDK6* (6q11), *ARID1B* (7q24) and *THADA* (2p21) had been translocated to *EVI1* in t(3;6), t(3;7) or t(2;3) AML, respectively (Figure 7B-C and Supplementary Fig. 6). All these genes are expressed in HSPCs [30]. Similar to the *MYC* SE in t(3;8) AML (Figure 4F), we found strong regulatory regions close to these, illustrated by H3K27ac and hematopoietic transcription factor binding (Figure 7B-D, Supplementary Fig. 6A). These commonalities suggest a shared mechanism for *EVI1* activation in all 3q26-rearranged leukemias, whereby an active hematopoietic enhancer is hijacked by a CTCF-mediated loop with the *EVI1* promoter. To validate this hypothesis, we targeted the *EVI1* CTCF binding site in MUTZ3-*EVI1*-eGFP, an inv(3) cell line engineered with eGFP as a reporter for *EVI1* [31]. In this model, mutations of the CTCF motif in a fraction of cells resulted in loss of eGFP and EVI1 expression (Supplementary Fig. 7A-C), as well as loss of CTCF binding (Supplementary Fig. 7D). Altogether, these results confirm the role of this CTCF binding site in enhancer hijacking leading to *EVI1* overexpression.

Figure 7. CTCF enhancer-docking site upstream of the *EVI1* promoter is preserved in 3q26-rearranged AML. **(**a) Schematic depiction of Chr.3 with a zoom-in on the *EVI1* locus, indicating the exact breakpoints (detected by 3q-seq) of 3q26-rearranged AMLs as vertical lines. In the lowest zoom-in panel the *EVI1* promoter with a CTCF binding site upstream marked respectively by H3K27ac (t(3;8) AML-17, orange) and CTCF (K562 t(3;8) clone 24-7, blue) ChIP-seq. (b) Schematic overview of Chr.6 and the locus where the breakpoints (black lines) were found by 3q-seq in t(3;6)(q26;q25) AML. The black box indicates the area of which the zoom-in is shown below. Zoom-in: putative enhancer indicated by H3K27ac (t(3;8) AML-17, orange), CTCF binding (K562 t(3;8) clone 24-7, blue) and HSPC active transcription factor recruitment (CD43+ cell [29], purple) at translocation site. The lines below indicate the exact breakpoints. The grey bar the minimal translocated region brought into close proximity of *EVI1* in that specific translocation. (c) Same as (b), but here for t(3;7)(q26;q11) AMLs. (d) Same as (b), but here for inv(3)/t(3;3)(q21;q26) AMLs. The exact translocated locus was previously shown to be an enhancer of *GATA2* [15].

DISCUSSION

In this study we investigated how *EVI1* is deregulated in AML with a translocation t(3;8)(q26;q24). Using an *EVI1-eGFP* t(3;8) model, we demonstrated that hyperactivation of *EVI1* was driven by a hijacked *MYC* SE. One enhancer module within this *MYC* SE, previously reported as enhancer module C [28], was particularly essential for *EVI1* transcription. Module C is reported to be responsible for *MYC* expression in primary leukemic cells. The high accessibility of this module and the binding of a core set of hematopoietic transcription factors drive *MYC* expression in HSPCs [28,29]. The other reported modules in the *MYC* SE, which did not affect *EVI1* transcription in a t(3;8) setting, may well be responsible for *MYC* transcription in other tissues [28]. Module C is the only element within the *MYC* SE to which early hematopoietic regulators bind, including GATA2, FLY1, ERG, RUNX1, LMO2 and LYL1 [29]. Since those factors also bind to other enhancers that recurrently translocate to *EVI1* in t(2;3)(p21;q26), t(3;7)(q26;q24), t(3;6)(q26;q11) or inv(3)/t(3;3)(3q26;3q21) AML, we argue that *EVI1* expression is driven by a common mechanism. This is in line with our previous published data on a variety of atypical 3q26-rearranged AMLs [20]. The loci donating their enhancer to *EVI1* harbor genes that are normally expressed in early HSPCs, e.g. *MYC, ARID1B, CDK6, THADA* or *GATA2.* Leukemias with high *EVI1* levels are chemotherapy-resistant and exhibit a unique gene expression signature comparable to that of CD34+ HSPCs [32]. This suggests that the cell of origin transformed in these leukemias is a very primitive hematopoietic progenitor cell.

The high-resolution 4C-seq data generated using our t(3;8) model revealed interaction of the *EVI1* promoter with the *MYC* SE, with multiple interaction zones associated with different enhancer modules indicative of a highly organized SE. Accordingly, Huang et al. defined the *MYC* SE as hierarchically organized. A hierarchical SE contains an enhancer module, referred to as hub enhancer, which is responsible for structural organization of the SE and is distinctly associated with CTCF and cohesin binding [7]. Module C was characterized as a hub enhancer within the *MYC* SE in K562 cells [7]. Interestingly, the deletion of module C, while leaving CTCF binding sites intact, not only affected *EVI1* expression but also disrupted *MYC* SE-*EVI1* promoter interaction. Furthermore, mutations in the vicinity of the CTCF core binding region also resulted in loss of interaction (Figure 6D). Altogether, this suggests that transcription factors and co-activators occupying this location play a role in enhancer-promoter interaction, either independently or in cooperation with CTCF. Analogous to CTCF, YY1 contributes to DNA-looping, but preferentially occupies interacting enhancers and promoters [8]. Although there is no indication that YY1 binds directly to enhancer module C, we did find YY1 binding flanking this module (Supplementary Fig. 8). In embryonic stem cells (ESC), pluripotency factors e.g. OCT4, NANOG and SOX2 recruit the mediator complex and stabilize the cohesin complex in order to facilitate cell type specific non-CTCF mediated enhancer-promoter looping [33]. In HSPCs a subunit of the mediator complex, MED12,

co-localizes with key hematopoietic transcription factors, interacting with additional transcriptional co-activators to maintain enhancer activity [34]. We hypothesize that in t(3;8) and other 3q26-rearranged AMLs, enhancer-promoter interaction is facilitated by CTCF and cohesin, which is further stabilized by recruitment of co-factors by hematopoietic regulators (see Graphical Abstract).

All the CTCF binding motifs in the *MYC* SE are oriented in a 'reverse' fashion, allowing a CTCF/cohesin complex to be formed with the 'forward' CTCF binding site 2.6 kb upstream of the *EVI1* transcriptional start site (TSS). In all 3q26-rearranged AMLs this upstream CTCF binding site was preserved with respect to *EVI1*. Interestingly, a CTCF binding site upstream of the *MYC* promoter has been reported to function as a docking site for enhancers driving *MYC* expression [6]. Our findings point to a very similar mechanism of transcriptional activation of *EVI1* in 3q26-rearranged AML. CTCF binding at this site proved to be absolutely critical for enhancer-promoter interaction and consequently indispensable for enhancer-driven *EVI1* transcription. Accordingly, it has been reported that promoters bound by CTCF, especially in enhancer deserts, are often dependent on long-range interactions [35].

Leukemias with 3q26 rearrangements depend on *EVI1*: interfering with EVI1 causes growth inhibition, differentiation and ultimately death of leukemic cells [15,31]. Our data demonstrate mechanistic similarities between the distinct enhancer-driven *EVI1*+ leukemias, suggesting that a therapy for one subtype may be effective for all these AMLs. The *EVI1-eGFP* t(3;8) model is a valuable tool for compound screens to identify inhibitors of *EVI1* transcription that could constitute a promising treatment for these refractory leukemias. As enhancer-driven transcription is not limited to leukemia, this model can also be used to study (super-) enhancer biology and transcriptional regulation in a broader context.

METHODS

All the materials and resources used in this study are summarized in Supplementary Table 1. This study did not generate any unique codes. All software tools used in this study are freely or commercially available and listed in Supplementary Table 1. All primer names and sequences can be found in Supplementary Data 1, which is an Excel file with multiple tabs listing sgRNAs, qPCR primers, amplicon seq primers, PCR primers and 4C primers.

Patient material

AML and T-ALL patient samples were collected either from the Erasmus MC Hematology department biobank (Rotterdam, The Netherlands) or from the MLL Munich Leukemia Laboratory biobank (Munich, Germany). Leukemic blast cells were purified from bone marrow or blood by standard diagnostic procedures. All patients provided written informed consent in accordance with the Declaration of Helsinki. The Medical Ethical Committee of the Erasmus MC has approved usage of the patient rest material for this study. The karyotype (gender) and age of the patients unique (when known) to this manuscript is given in the Source data file.

Generation of *EVI1* expression cell model

The plasmids to clone T2A-eGFP in frame with *EVI1* were designed and described by my colleagues as follows [31]. The repair template was generated using Gibson Assembly (NEB). Both homology arms were PCR amplified from MUTZ3 genomic DNA using Q5 polymerase (NEB). The first homology arm consists of a part of the intron and last exon of *EVI1* minus the STOP, the second homology arm consists of part of the 3'UTR with the PAM sequence of sgRNA omitted. The *T2A-eGFP* was PCR amplified from dCAS9-VP64_2A_GFP. All fragments were cloned using the Gibson assembly into the PUC19 backbone. The sgRNA sequence AGCCACGTATGACGTTATCA was cloned into pX330-U6-Chimeric_BB-CBh-hSpCas9. Cells were nucleofected with pX330 vector containing the sgRNA and Cas9 and the repair template using the NEON transfection system (Thermo Fisher Scientific) with buffer R and program 3 (1350 V, 10 ms, 4 pulses). GFP^+ cells were sorted using a FACS AriaIII (BD Biosciences), and after two rounds of enrichment for cells expressing eGFP+, these cells were single cell sorted and tested for proper integration. Subsequently, clones were named K562 *EVI1-eGFP*; multiple clones were obtained, but in this study only clone 8 and 24 were used for further experiments.

Generation of a t(3;8)(q24;q26) model

K562 *EVI1-eGFP* clones (clone 8 and clone 24) were used as parental clones to generate the t(3;8)(q24;q26) clones. Based on the breakpoints (Chr.3:168.917.999 - Chr.8:130.487.191) of primary AML sample (#HF-80), sgRNAs were designed (using ChopChop V3 [36], Supplementary

Data 1) and mixed with purified Cas9 (IDT) to make ribonucleoproteins (RNPs). The NEON transfection system (Thermo Fisher Scientific) was used to get the RNPs into the K562 *EVI1-eGFP* clones. Three days after transfection the eGFP+ cells were sorted using the FACS AriaIII, and this enrichment process was repeated twice before eGFP+ single cells were sorted to produce single cell clones. The clones were characterized for the designed specific t(3;8)(q24;q26) translocation by PCR (primers in Supplementary Data 1), Sanger-seq, flow cytometry and FISH.

Cytogenetics: karyotype and FISH

Diagnostic cytogenetics for all samples was performed by each of the institutes mentioned above. For this study, samples were selected based on t(3;8)(q26;q24) rearrangements detected by karyotyping and/or *MECOM* interphase fluorescence *in situ* hybridization (FISH). FISH and classic metaphase karyotyping were performed and reported according to standard protocols based on the International System of Human Cytogenetics Nomenclature (ISCN) 2017 [37]. For both patient samples and K562 clones *MECOM* FISH was performed according to the manufacturer's protocol, using the *MECOM* t(3;3); inv(3)(3q26) triple color probe (blue, green, red, Cytocell, LPH-036). For the characterization of the K562 *EVI1-eGFP* t(3;8) clones the *MECOM* FISH was combined with: CEP8 (cen.8, blue), IGH (14q32, green), C-MYC(8q24, orange) (Vysis, 04N10-020) and C8 (Vysis, SpO, 07J22-008).

Targeted chromosomal region 3q21.1-3q26.2 DNA sequencing (3q-seq)

Genomic DNA was fragmented with the Covaris shearing device (Covaris), and sample libraries were constructed with the KAPA Hyper Prep Kit (Roche). After ligation of adapters and an amplification step, target sequences of chromosomal regions 3q21.1-q26.2 were captured by using custom in-solution oligonucleotide baits (Nimblegen SeqCap EZ Choice XL). Amplified captured sample libraries were paired-end sequenced (2x100 bp) on the HiSeq 2500 platform (Illumina) and aligned against the Human Genome Assembly 19 (hg19) using the Burrows-Wheeler aligner[38] v0.7.17. All chromosomal aberrations, such as translocations and inversions, were determined with BreakDancer v1.1 [39].

RNA isolation, quantitative PCR (qPCR) and RNA sequencing

RNA was isolated using TRIzol (Invitrogen) or the Allprep DNA/RNA mini kit (Qiagen). cDNA was synthesized using SuperScript II Reverse transcriptase (Invitrogen). Quantitative real-time RT-PCR was performed on the 7500 Fast Real time PCR System (Thermo Fisher Scientific) with 10 ul Fast Sybr Green Master Mix (Thermo Fisher Scientific), 2 ul of cDNA and primers listed in Supplementary Data 1. Relative levels of gene expression were calculated using the ΔΔ Ct method [40]. For qPCR data one-way ANOVA (GraphPad PRISM) was performed

to indicate level of significant differences between clones or conditions. For qPCR data of cells directly after FACS no statistical test could be performed due to the limited number of cells (Figure 4 C-D). RNA-seq data from non-3q26 AMLs and CD34+ have been previously published in [41] and are accessible at the European Genome-phenome Archive (EGA) under accession number EGAS00001004684.

Sample libraries were prepared using 500 ng of input RNA according to the KAPA RNA HyperPrep Kit with RiboErase (HMR) (Roche) using Unique Dual Index adapters (Integrated DNA Technologies, Inc.). Amplified sample libraries were paired-end sequenced (2x100 bp) on the Novaseq 6000 platform (Illumina). Salmon[42] v0.13.1 was used to quantify expression of individual transcripts, which were subsequently aggregated to estimate gene-level abundances with tximport[43]. Differential gene expression analysis of count estimates from Salmon was performed with DEseq2[44]. The results of this analysis were depicted as a boxplot using the ggplot2 R package.

Cell lines and culture

K562 cells were cultured in RPMI 1640 + L-glutamine (Hyclone SH30027.LS), 10% fetal calf serum (FCS, Gibco) and 50 U/mL penicillin and 50 µg/mL streptomycin (Gibco 15140-163). Cells were incubated at 37°C and 5% CO_2 and passaged every 3-4 days to 100.000 cells/ml. A previously generated MUTZ3 *EVI1*-eGFP cell line was cultured in αMEM (HyClone) with 20% fetal calf serum (FCS, Gibco) and 20% conditioned 5637 medium [31]. Unique biological materials are available upon request by contacting the corresponding author.

Genome editing

CRISPR-Cas9 technology was used to make mutations or deletions in the regions described in the results section. All primer sequences to generate sgRNAs can be found in Supplementary Data 1 and were ordered from IDT. By in vitro transcription sgRNAs were produced as described above for the generation of the t(3;8). In short: the constant and specific oligos were annealed and filled in 20 min 12°C by T4 polymerase (NEB, M0203S), sgRNAs were produced by in vitro transcription using HIScribe T7 High-Yield RNA Synthesis kit (NEB, E2050S) 3-4h, 37°C, DNA was eliminated by Turbo DNase (Thermo Fisher Scientific, AM2238), 15min, 37°C. The sgRNas were concentrated and purified using RNA clean & concentrator -25 (Zymo Research, R1017). The concentration of sgRNAs was estimated using Qubit RNA BR assay (Invitrogen, Q10210). Ribonucleoproteins (RNPs) were made by mixing purified Cas9 protein (IDT, Nucleofection of all K562 clones was done with NEON transfection buffer T (Thermo Fisher Scientific) and settings 1350V, 10ms, 4 pulses. Nucleofection of MUTZ3 *EVI1*-eGFP cells was done with NEON transfection buffer R (Thermo Fisher Scientific) and settings 1500V, 20ms, 1 pulse. After a minimum of 72 hrs post nucleofection DNA or RNA was extracted (DNA Quick extract, Epicenter or Qiagen Allprep DNA/RNA mini, #80204) or cells were harvested for further analysis by respectively PCR, qPCR or flow cytometry analysis/FACS sorting.

The targeted CTCF motifs in the *EVI1* promoter were identified using the CTCFBSDB 2.0 database[45]. CTCF motif orientation at the *EVI1* promoter and the *MYC* SE was retrieved from the JASPAR database (release 2020) [46].

Flow cytometry and sorting (FACS)

Flow cytometric analysis or cell sorting was performed using the FACS Canto or the FACS Aria flow cytometer (BD Biosciences). Cells were gated on viability and single cells using FSC/SSC, eGFP intensity levels were measured using the FITC channel. Data were analyzed using FACS Diva v9.0 and FlowJo v10.0.

PCR and primers

For all PCRs used to detect translocations, point mutations or deletions; Q5 High-Fidelity DNA polymerase was used following the manufactures protocol (NEB, #M0491) and primers listed in Supplementary Data 1. PCR products were purified using a Qiaquick PCR purification kit. Purified PCR products were subjected to Sanger sequencing on an Applied Biosystems 3730 device using a BigDye™ Terminator v1.1 Cycle Sequencing Kit (Thermo Fisher Scientific) and primers listed in Supplementary Data 1.

Amplicon sequencing

To check mutations after targeting with CRISPR-Cas9 we performed amplicon sequencing using the Illumina PCR-based custom amplicon sequencing method using the TruSeq Custom Amplicon index kit (Illumina). The first PCR was performed using Q5 polymerase (NEB), the second nested PCR with KAPA HiFi HotStart Ready mix (Roche). Samples were sequenced paired-end (2x 250bp) on a MiSeq (Illumina). Reads were trimmed with trimgalore[47] v0.4.4 to remove low-quality bases and adapters, and subsequently aligned to the human reference genome build hg19 with BBMap[48] v34.92 allowing for 1000 bp indels. Mutations introduced by genome editing were analyzed and visualized using CRISPResso2[49] v2.0.27.

Western Blotting

Cells were lysed using the NE-PER Nuclear and Cytoplasmic Extraction Kit (Thermo Fisher Scientific) following the manufacturer' protocol and nuclear extract was used for Western Blotting of EVI1 (#2265 Cell Signaling, dilution: 1:1000). As loading control an antibody against B-Actin (clone AC15, A5441, Sigma, dilution: 1:10.000) was used. The Odyssey infrared imaging system (Li-Cor) was used for visualization of the protein levels.

4C sequencing

Chromosome Conformation Capture Sequencing (4C-seq) sample preparation was performed using 10 million cells [50]. In short, genomic regions that are spatially proximal in the cell nucleus were fixated by formaldehyde-induced crosslinks. The DNA was fragmented with DpnII as a primary restriction enzyme, Csp6I as a secondary 4 bp-cutter. To identify and

quantify fragments that were ligated to the genomic region of interest, a two-step PCR was performed [51]. The first PCR step was an inverse PCR with viewpoint-specific primers that are listed in Supplementary Data 1. In the second PCR step, universal primers were used that contain the Illumina adapters. The amplicons were subjected to next generation sequencing on the Illumina NovaSeq 6000 platform.

Demultiplexing and clipping of the primer sequences was performed by an in-house algorithm. Subsequently, the reads of each viewpoint were aligned against the human genome (hg19) with bowtie[52] v1.1.1. Reads not mapping to fragments determined by the restriction site positions of the chosen primary and secondary restriction enzymes were removed by an in-house algorithm. Generated BAM-files were transformed into WIG-files with an in-house tool, applying a running mean (window size 21) for signal smoothing of peaks. The data were also normalized to reads per million (RPM). In all figures, the tracks were displayed on the Integrative Genomics Viewer (IGV)[53] v2.8 using "group auto-scale" to compare relevant samples.

ChIP sequencing

Chromatin immunoprecipitation sequencing (ChIP-seq) experiments were performed using 10 to 20 million cells. Cells were cross-linked with 1% formaldehyde. Chromatin was isolated using lysis buffer A (50mM Tris pH8, 10mM EDTA, 1% SDS). In the CTCF ChIP, 0.5% EPIGEN BB was added to the lysis buffer A. In the RUNX1 ChIP at least 30 million cells were double crosslinked with 2mM disuccinimidyl glutarate followed by 1% formaldehyde. Chromatin of double crosslinked cells was isolated using lysisbuffer B (10mM Tris pH7.5, 74mM NaCl, 3mM MgCl2, 1mM CaCl2, 4% NP40, 0.32% SDS). The chromatin was sonicated with a Bioruptor device (Diagenode) using the following settings: 10 cycles of 30 sec on, 30 sec off.

Immunoprecipitation of cross-linked chromatin was performed with antibodies directed against H3K27Ac (Diagenode C15410196, 2.5 ug), H3K9Ac (Diagenode C15410004, 2.5 ug), H3K4me3 (Diagenode C15410003, 2.5 ug), RUNX1 (Abcam Ab23980, 5 ug) in IP dilution buffer (16.7mM Tris pH8, 1.2 mM EDTA, 167mM NaCl, 1.1% Triton, 0.01% SDS) or CTCF (Cell Signalling 2899S, 5 ug) in CTCF IP dilution buffer (20mM Tris, 2mM EDTA, 100mM NaCl, 0.5% Triton). Chromatin bound antibody was precipitated with prot G Dynabeads (Thermo Fisher Scientific) and washed with low salt buffer (20mM Tris pH8, 2mM EDTA, 1% Triton, 150mM NaCl), high salt buffer (20mM Tris pH8, 2mM EDTA, 1% Triton, 500mM NaCl), LiCl buffer (10mM Tris, 1mM EDTA, 0.25mM LiCl, 0.5% IGEPAL, 0.5% Sodium-Deoxycholate) and TE (10mM Tris pH8, 1mM EDTA). Chromatin was eluted in elutionbuffer A (25mM Tris, 10mM EDTA, 0.5% SDS). In the CTCF ChIP and RUNX1 ChIP chromatin was eluted in elution buffer B (0.1M Sodiumhydrogencarbonate, 1% SDS).

Crosslinks were reversed overnight at 65°C in the presence of proteinase K (New England Biolabs). De-crosslinked material was purified using a QIAGEN PCR Purification Kit. The purified DNA was processed according to the Nextflex ChIP Sample Preparation

Protocol (Perkin Elmer) or the Microplex library preparation kit V2 (Diagnode C05010013) and sequenced on the Illumina NovaSeq6000 platform. ChIP-seq reads were aligned to the human reference genome build hg19 with bowtie[52] v1.1.1 and bigwig files were generated for visualization with the bamCoverage tool from deepTools[54] v3.4.3, with the options --normalizeUsing RPKM --smoothLength 100 --binSize 20. Peak calling was performed with MACS2[55] v2.2.7.1 using default settings. Publicly available ChIP-seq data were downloaded from the Gene Expression Omnibus: RAD21 and SMC3 tracks in K562 cells generated by the ENCODE consortium[56], MED12 data in K562 published by the Aifantis group[57], and hematopoietic transcription factors in CD34+ cells generated by the Pimanda group[29]. In all figures displaying ChIP-seq data the y-axis shows normalized RPKM, and "group auto-scale" was used on IGV[53] v2.8 to compare relevant samples.

ATAC sequencing

Cells were washed using PBS and counted in a Bürker-Türk counting chamber. 50.000-100.000 cells were pelleted and resuspended in 1 ml ATAC lysing buffer containing : 0.3 M Sucrose, 10 mM Tris HCL pH 7.5, 60 mM KCl, 15 mM NaCL, 5 mM MgCl, 0.1 mM EGTA, 0.1% NP40, 0.15 mM Spermine, 0.5 mM Spermidine and 2 mM 6AA. All components were derived from Sigma Aldrich. Cells were incubated in lysis buffer for 3 minutes on ice. Cells were pelleted at 500xg for 10 minutes at 4°C and supernatant removed. Pelleted cells were resuspended in 50 µl transposase mixture containing 25 µl 2x TD buffer, 2.5 µL TD1 transposase, 22.5 µl nuclease free water (kit Illumina cat no 20034197). Samples were incubated 30 minutes at 37°C while mixing at 500 RPM in a heat block. Samples were immediately purified using the Qiagen min elute PCR purification kit following manufacturers protocol. Transposase fragmented DNA was eluted in 10 µl elution buffer. All DNA was used in a 4 cycle PCR amplification using Nextera i7- and i5-index primers (Illumina). 5 µl of the 4 cycle amplified material was used in taqman. ¼ of the maximum signal was determined and cycles were added to the remaining 45 µl library to avoid over-amplification of the ATAC library. Amplified libraries were purified using Agencourt AMPure XP beads using a 1;1,8 ratio. DNA was eluted using 30 µl EB buffer. Libraries were quantified using Qubit and PCR NEBnext library quant kit for Illumina (NEB). Size distribution was determined by running 1 ng library on a DNA high sensitive chip (Agilent / Bioanalyzer).

ATAC-seq samples were sequenced paired-end 2x50 bp or 2x100 bp on the Hiseq 2500 and the Novaseq 6000 platforms (both Illumina). They were aligned against the human genome (hg19) with bowtie2[58] v2.3.4.1, allowing for a maximum 2000 bp insert size. Mitochondrial reads and fragments with mapping quality below 10 were removed. bigwig files were generated for visualization with the bamCoverage tool from deepTools v3.4.3 [54], with the options --normalizeUsing RPKM --smoothLength 100 --binSize 20. In all figures displaying ATAC-seq data the y-axis shows normalized RPKM, and "group auto-scale" was used on IGV [53] v2.8 to compare relevant samples.

Comparative analysis of modules in the MYC super-enhancer

Quantification of H3K27ac and ATAC-seq reads was conducted in the different enhancer modules within the MYC super-enhancer, as defined in [28]. A BED file containing these modules was converted into GTF with the UCSC tools bedToGenePred and genePredToGtf[59]. Read counts in enhancer regions were computed with featureCounts[60] and differential analysis was conducted with the DESeq2 R package[44]. The results of this analysis were depicted as a boxplot using the ggplot2 package in R.

SNP array

DNA was isolated from K562 cells using the AllPrep DNA/RNA mini kit (Qiagen, #80204). All SNP arrays were performed at the Erasmus MC Department of Clinical Genetics (Rotterdam, The Netherlands) and analyzed as previously described [20,61,62]. In short, 200 ng of DNA was used as an input for a single array. DNA amplification, tagging and hybridisation were performed according to the manufacturer's protocol. The array slides were scanned on an iScan Reader (Illumina). Data analysis was performed using GenomeStudio version 2.0, KaryoStudio version 1.4 (Illumina, standard settings) and Nexus Copy Number 9.0 (BioDiscovery, El Segundo, CA, USA).

Statistics and Reproducibility

The *EVI1* knockdown experiment shown in Figure 2c and Supplementary Fig. 1d were performed in 2 biological replicas; in clone 8 and clone 24. The *EVI1* knockdown experiment shown in Figure 2l and Supplementary Fig. 2f were performed in 3 biological replicas; in clone 8-4 and clone 24-7 and clone 24-12. The PCR over the t(3;8) breakpoint to identify single clones that harbored the translocation, shown in Figure 2h, was done on over 20 single clones/biological replicas. Uncropped PCR gel pictures are provided in the Source Data file. The deletions induced in the *MYC* super-enhancer as shown by PCR in Figure 4b (right panel) were performed in 3 biological replicas of which 2 t(3;8) clones: clone 8-4 and 24-7 and one control clone harboring a 3q/*MECOM* amplification: clone 24-2 (as characterized in Supplementary Fig. 4). The CTCF binding site deletions induced in the *MYC* SE as shown by PCR in Figure 5b (right panel) and the single cut by shRNA#1 were performed as minimum with 2 biological replicas in t(3;8) clone 8-4 and the control clone 24-2. However, most important experiments like deletion of enhancer module C, the deletion CTCF2 in the *MYC* SE of CTCF near the *EVI1* promoter of the single cut in the CTCF site at the *EVI1* promoter (sgRNA#1) were performed in 3 biological replicas in clone 8-4, 24-7 and control clone 24-2 and at least twice in the two separate t(3;8) clones. Uncut PCR gel pictures are provided in the Source data file. The western blot showing the EVI1 protein levels after sorting in Figure 4e was performed in 2 biological replicas. Uncut blot pictures are provided in the Source data file. Genotyping of the K562 *EVI1-eGFP* clones (Supplementary Fig. 1b and 1c) was done for 10 single clones. Clones with the correct genotype were selected based on at least 3 different PCR methods. Uncut PCR gel pictures are provided in the Source data file.

The FISH experiments were done by the Erasmus MC diagnostic lab following their verified experimental set up, the FISH experiments as shown in Supplementary Fig. 2c, 2d, 2e and 4d were done on 4 separate clones each. All with similar results as shown in the pictures in the main manuscript. The PCR on the sorted fractions was performed in 3 biological replicas in 2 separate t(3;8) clones (1x clone 8-4 and 2x 24-7). Uncut PCR gel pictures are provided in the Source data file. In Supplementary Fig. 4 the control clone 24-4 harboring the 3q/*MECOM* amplification is characterized. In total we generated 4 clones (4 biological replicas) like this of which two clones are shown (24-1 and 24-2) in Supplementary Fig. 4b. The uncut PCR gel picture showing all 4 clones is provided in the Source data file. The deletions induced by CRISPR-Cas9 as shown in Supplementary Fig. 4g are done all at least in 2 biological replicas. This control clone 24-2 was always taken along in CRISPR-CAs9 experiments as a (negative) control for an effect on EVI1/eGFP expression.

DATA AVAILABILITY

The ChIP-seq, 3q-seq, 4C-seq and RNA-seq data derived from human patients are available at the European Genome-phenome Archive (EGA), under the accession code EGAS00001004808 [https://ega-archive.org/studies/EGAS00001004808]. These data are available under restricted access due to data privacy laws, access can be obtained by contacting the Data Access Committee and signing a Data Access Agreement.

Data derived from K562 have been uploaded to the ArrayExpress database under the following accession codes: E-MTAB-9958 (4C-seq) [https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-9958/], E-MTAB-9965 (ChIP-seq) [https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-9965/], E-MTAB-10785 (ATAC-seq) [https://www.ebi.ac.uk/arrayexpress/experiments/MTAB-10785/] and E-MTAB-9937 [https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-9965/], (Amplicon-sequencing following CRISPR-Cas9 treatment).

This study also used publicly available sequencing datasets. The 3q-seq data of the inv(3)/t(3;3) cell lines MUTZ3 and MOLM1 were downloaded from ArrayExpress, under the accession code E-MTAB-2224 [https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2224/]. The ChIP-seq data of a heptad of transcription factors in CD34+ cells generated by the Pimanda group [29] were downloaded from the Gene Expression Omnibus (GEO), under the accession code GSE38865 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE38865]. The ChIP-seq data of RAD21, SMC3 and YY1 generated by the ENCODE consortium [56] were also downloaded from GEO, under the accession code GSE31477 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31477]. The RNA-seq data of HSPCs generated by the Blueprint consortium [30] was accessed via the Blueprint Data Analysis Portal [http://blueprint-data.bsc.es/release_2016-08/]

Source data is provided with this paper. All uncut blots and gel pictures can be found in the Excel Source file.

AUTHOR CONTRIBUTIONS

S.O., L.S and R.D. conceived of and designed the study. B.W., C.H. and T.H. provided study materials or patient samples. H.B.B supervised all cytogenetic and FISH characterization. C.E-V., S.v.H., M.H., A.A.V., L.S., S.G. and S.O. performed experiments, data analysis and interpretation. M.V. operated the FACS sorter for a significant part of the experiments. E.B. organized all NGS and was involved in data interpretation. R.M-L. was responsible for all bioinformatics data processing and analysis of this study. S.O., R.M-L., L.S. and R.D. wrote the manuscript.

My contributions to this work were: processing and analysis of all high throughput sequencing data (3q-capture, RNA-seq, 4c-seq, amplicon-seq, ATAC-seq, ChIP-seq); data management and upload; interpretation of the results and writing of the manuscript.

ACKNOWLEDGEMENTS

The authors are indebted to their colleagues from the bone marrow transplantation group and the molecular and cytogenetics diagnostics laboratories of the Department of Hematology and Clinical Genetics at Erasmus University Medical Center for storage of samples and molecular analysis of the leukemia cells (M. Wattel, R. van der Helm and P.J.M. Valk). For providing patient material, the authors are thankful to the MLL Münchner Leukämielabor GmbH in Germany. They also thank P. Sonneveld and their colleagues of the Hematology Department, especially those involved in FACS sorting (C. van Dijk), Next Generation Sequencing operating, bioinformatics (R. Hoogenboezem) and all others for their input or expertise. We also thank N.J. Galjart and R. Stadhouders of the department of Cell Biology and Pulmonary medicine at the Erasmus University Medical Center for their input and expertise.

This work was funded by grants and fellowships from the Dutch Cancer Society (R.D., R.M-L., S.O., L.S.), Skyline DX (S.O.), the Daniel den Hoed, Erasmus MC Foundation (L.S.).

CONFLICT OF INTEREST DISCLOSURE

T.H. and C.H. are employees of and have equity ownership in MLL Munich Leukemia Laboratory. The remaining authors declare no competing interests.

REFERENCES

1 Stadhouders, R., Filion, G. J. & Graf, T. Transcription factors and 3D genome conformation in cell-fate decisions. *Nature* **569**, 345-354 (2019).

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

3 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).

4 Dowen, J. M. *et al.* Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374-387 (2014).

5 Zuin, J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci U S A* **111**, 996-1001 (2014).

6 Schuijers, J. *et al.* Transcriptional Dysregulation of MYC Reveals Common Enhancer-Docking Mechanism. *Cell Rep* **23**, 349-360 (2018).

7 Huang, J. *et al.* Dissecting super-enhancer hierarchy based on chromatin interactions. *Nat Commun* **9**, 943 (2018).

8 Weintraub, A. S. *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* **171**, 1573-1588 e1528 (2017).

9 Bulger, M. & Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327-339 (2011).

10 Spitz, F. Gene regulation at a distance: From remote enhancers to 3D regulatory ensembles. *Semin Cell Dev Biol* **57**, 57-67 (2016).

11 Muerdter, F. & Stark, A. Gene Regulation: Activation through Space. *Curr Biol* **26**, R895-R898 (2016).

12 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674 (2011).

13 Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer* **7**, 233-245 (2007).

14 Yamazaki, H. *et al.* A remote GATA2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating EVI1 expression. *Cancer Cell* **25**, 415-427, doi:10.1016/j.ccr.2014.02.008 (2014).

15 Groschel, S. *et al.* A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* **157**, 369-381 (2014).

16 Lugthart, S. *et al.* High EVI1 levels predict adverse outcome in acute myeloid leukemia: prevalence of EVI1 overexpression and chromosome 3q26 abnormalities underestimated. *Blood* **111**, 4329-4337, doi:10.1182/blood-2007-10-119230 (2008).

17 Gröschel, S. *et al.* High EVI1 Expression Predicts Outcome in Younger Adult Patients With Acute Myeloid Leukemia and Is Associated With Distinct Cytogenetic Abnormalities. *Journal of Clinical Oncology* **28**, 2101-2107, doi:10.1200/jco.2009.26.0646 (2010).

18 Lugthart, S. *et al.* Clinical, Molecular, and Prognostic Significance of WHO Type inv(3)(q21q26.2)/t(3;3)(q21;q26.2) and Various Other 3q Abnormalities in Acute Myeloid Leukemia. *Journal of Clinical Oncology* **28**, 3890-3898, doi:10.1200/jco.2010.29.2771 (2010).

19 Barjesteh van Waalwijk van Doorn-Khosrovani, S. *et al.* High EVI1 expression predicts poor survival in acute myeloid leukemia: a study of 319 de novo AML patients. *Blood* **101**, 837-845, doi:10.1182/blood-2002-05-1459 (2003).

20 Ottema, S. *et al.* Atypical 3q26/MECOM rearrangements genocopy inv(3)/t(3;3) in acute myeloid leukemia. *Blood* **136**, 224-234 (2020).

21 Lin, P., Medeiros, L. J., Yin, C. C. & Abruzzo, L. V. Translocation (3;8)(q26;q24): a recurrent chromosomal abnormality in myelodysplastic syndrome and acute myeloid leukemia. *Cancer Genetics and Cytogenetics* **166**, 82-85, doi:https://doi.org/10.1016/j.cancergencyto.2005.10.007 (2006).

22 Lennon, P. A. *et al.* Aberrant EVI1 expression in acute myeloid leukemias associated with the t(3;8)(q26;q24). *Cancer Genetics and Cytogenetics* **177**, 37-42, doi:https://doi.org/10.1016/j.cancergencyto.2007.05.007 (2007).

23 De Braekeleer, M. *et al.* Breakpoint heterogeneity in (2;3)(p15–23;q26) translocations involving EVI1 in myeloid hemopathies. *Blood Cells, Molecules, and Diseases* **54**, 160-163, doi:https://doi.org/10.1016/j.bcmd.2014.11.015 (2015).

24 Trubia, M. *et al.* Characterization of a recurrent translocation t(2;3)(p15–22;q26) occurring in acute myeloid leukaemia. *Leukemia* **20**, 48-54, doi:10.1038/sj.leu.2404020 (2006).

25 Storlazzi, C. T. *et al.* A novel chromosomal translocation t(3;7)(q26;q21) in myeloid leukemia resulting in overexpression of EVI1. *Annals of Hematology* **83**, 78-83, doi:10.1007/s00277-003-0778-y (2004).

26 Nucifora, G., Laricchia-Robbio, L. & Senyuk, V. EVI1 and hematopoietic disorders: History and perspectives. *Gene* **368**, 1-11, doi:https://doi.org/10.1016/j.gene.2005.09.020 (2006).

27 Tang, G. *et al.* t(3;8)(q26.2;q24) Often Leads to MECOM/MYC Rearrangement and Is Commonly Associated with Therapy-Related Myeloid Neoplasms and/or Disease Progression. *J Mol Diagn* **21**, 343-351 (2019).

28 Bahr, C. *et al.* A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. *Nature* **553**, 515-520 (2018).

29 Beck, D. *et al.* Genome-wide analysis of transcriptional regulators in human HSPCs reveals a densely interconnected network of coding and noncoding genes. *Blood* **122**, e12-22 (2013).

30 Fernandez, J. M. *et al.* The BLUEPRINT Data Analysis Portal. *Cell Syst* **3**, 491-495 e495, doi:10.1016/j.cels.2016.10.021 (2016).

31 Smeenk, L. *et al.* Selective requirement of MYB for oncogenic hyperactivation of a translocated enhancer in leukemia. *Cancer Discov*, doi:10.1158/2159-8290.CD-20-1793 (2021).

32 Valk, P. J. M. *et al.* Prognostically Useful Gene-Expression Profiles in Acute Myeloid Leukemia. *New England Journal of Medicine* **350**, 1617-1628, doi:10.1056/NEJMoa040465 (2004).

33 Denholtz, M. & Plath, K. Pluripotency in 3D: genome organization in pluripotent cells. *Curr Opin Cell Biol* **24**, 793-801 (2012).

34 Aranda-Orgilles, B. *et al.* MED12 Regulates HSC-Specific Enhancers Independently of Mediator Kinase Activity to Control Hematopoiesis. *Cell Stem Cell* **19**, 784-799 (2016).

35 Kubo, N. *et al.* Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation. *Nat Struct Mol Biol* **28**, 152-161, doi:10.1038/s41594-020-00539-5 (2021).

36 Labun, K. *et al.* CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Research* **47**, W171-W174, doi:10.1093/nar/gkz365 (2019).

37 International Standing Committee on Human Cytogenomic, N., McGowan-Jordan, J., Simons, A. & Schmid, M. *ISCN : an international system for human cytogenomic nomenclature (2016)*. (2016).

38 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

39 Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**, 677-681, doi:10.1038/nmeth.1363 (2009).

40 Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402-408, doi:10.1006/meth.2001.1262 (2001).

41 Mulet-Lazaro, R. *et al.* Allele-specific expression of GATA2 due to epigenetic dysregulation in CEBPA double mutant AML. *Blood*, doi:10.1182/blood.2020009244 (2021).

42 Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419, doi:10.1038/nmeth.4197 (2017).

43 Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521, doi:10.12688/f1000research.7563.2 (2015).

44 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

45 Ziebarth, J. D., Bhattacharya, A. & Cui, Y. CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization. *Nucleic Acids Res* **41**, D188-194, doi:10.1093/nar/gks1165 (2013).

46 Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **48**, D87-D92, doi:10.1093/nar/gkz1001 (2020).

47 Trim Galore (2012).

48 Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner. (2014).

49 Clement, K. *et al.* CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat Biotechnol* **37**, 224-226, doi:10.1038/s41587-019-0032-3

10.1038/s41587-019-0032-3 [pii] (2019).

50 van de Werken, H. J. *et al.* Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat Methods* **9**, 969-972, doi:10.1038/nmeth.2173 (2012).

51 Krijger, P. H. L., Geeven, G., Bianchi, V., Hilvering, C. R. E. & de Laat, W. 4C-seq from beginning to end: A detailed protocol for sample preparation and data analysis. *Methods* **170**, 17-32, doi:10.1016/j.ymeth.2019.07.014 (2020).

52 Langmead, B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics* **Chapter 11**, Unit 11 17, doi:10.1002/0471250953.bi1107s32 (2010).

53 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).

54 Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165, doi:10.1093/nar/gkw257 (2016).

55 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137-R137, doi:10.1186/gb-2008-9-9-r137 (2008).

56 Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

57 Aranda-Orgilles, B. *et al.* MED12 Regulates HSC-Specific Enhancers Independently of Mediator Kinase Activity to Control Hematopoiesis. *Cell Stem Cell* **19**, 784-799, doi:10.1016/j.stem.2016.08.004 (2016).

58 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

59 Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief Bioinform* **14**, 144-161, doi:10.1093/bib/bbs038 (2013).

60 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2014).

61 Srebniak, M. *et al.* Application of SNP array for rapid prenatal diagnosis: implementation, genetic counselling and diagnostic flow. *European Journal Of Human Genetics* **19**, 1230, doi:10.1038/ejhg.2011.119 (2011).

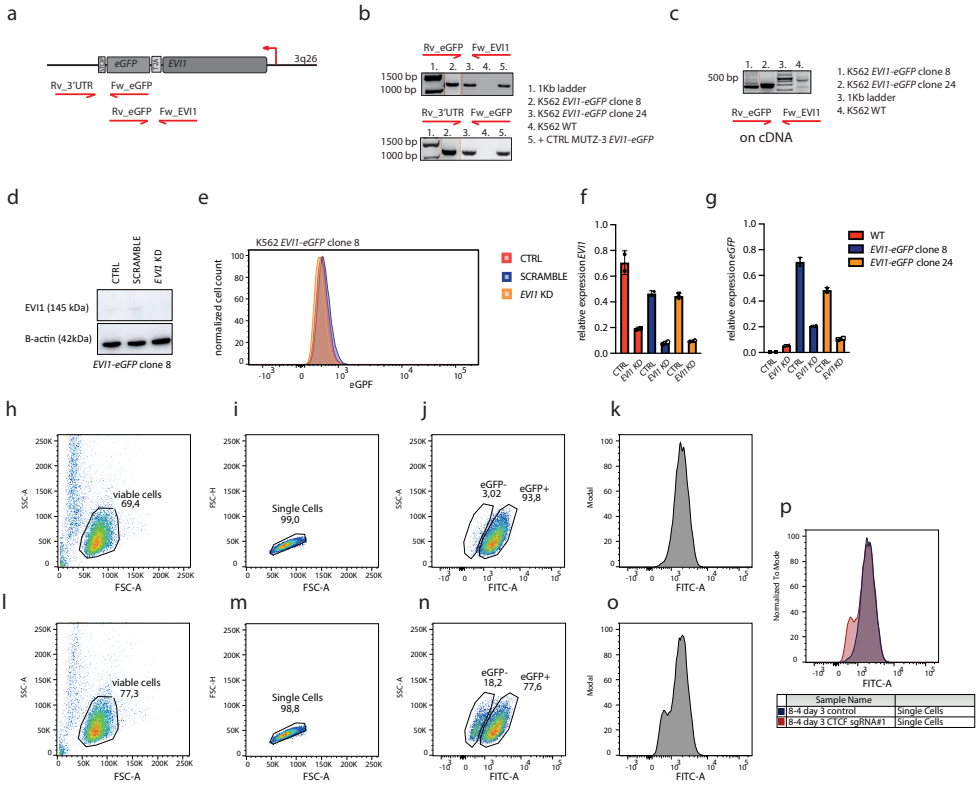
62 Srebniak, M. I. *et al.* Prenatal SNP array testing in 1000 fetuses with ultrasound anomalies: causative, unexpected and susceptibility CNVs. *European Journal Of Human Genetics* **24**, 645, doi:10.1038/ejhg.2015.193 (2015).

63 Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* **46**, D794-D801 (2018).

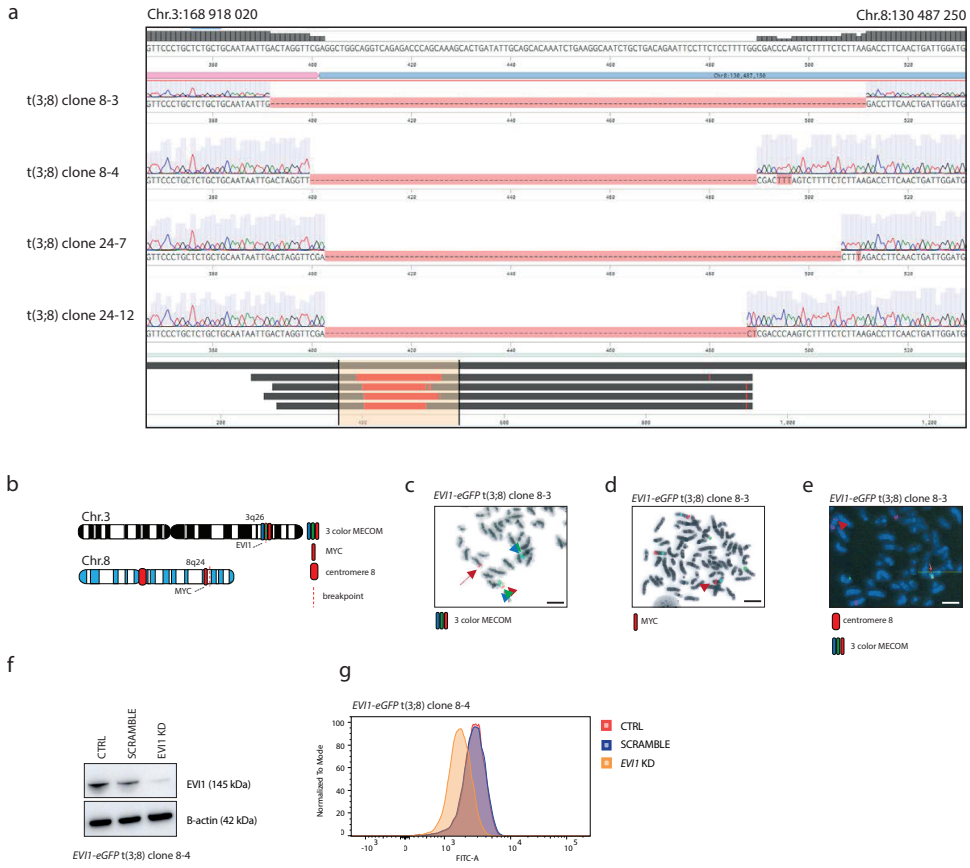
64 Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **48**, D87-D92, doi:10.1093/nar/gkz1001 (2019).

SUPPLEMENTARY INFORMATION

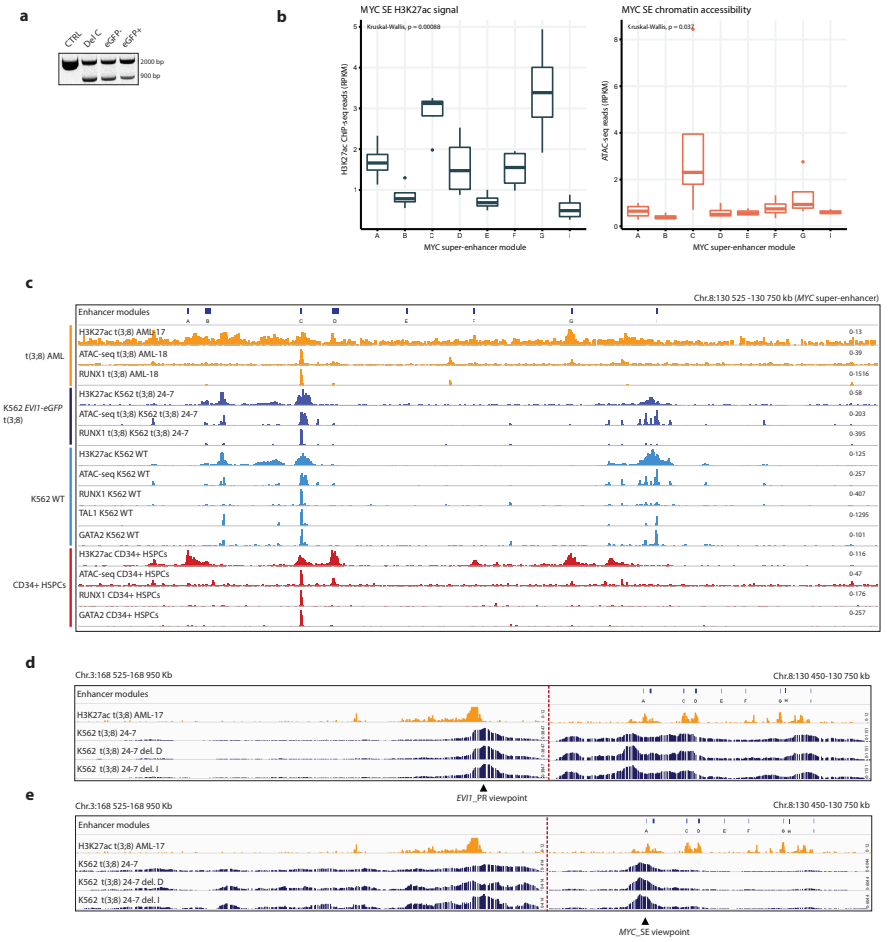
SUPPLEMENTARY FIGURES



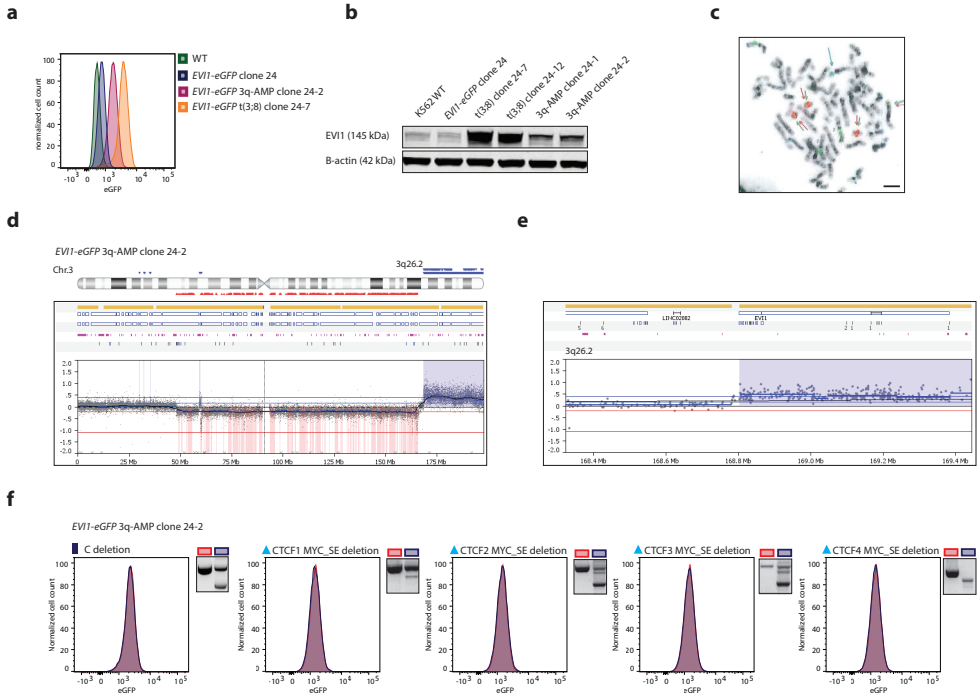
Supplementary Figure 1. Generation of cell model with eGFP reporter for *EVI1* expression. (a) Schematic overview of PCR primers located in the *eGFP-T2A-EVI1* insert. (b) PCRs of genomic DNA of K562 *EVI1-eGFP* clones to verify correct genotype. Source data are provided as a Source Data file. (c) PCR on cDNA to verify *EVI1-eGFP* transcript in the same clones. Source data are provided as a Source Data file. (d) Western blot showing very low EVI1 protein levels in the controls as expected, and absent EVI1 protein levels upon *EVI1* knockdown by shRNA. Source data are provided as a Source Data file. (e) Flow cytometry plot showing eGFP reduction upon *EVI1* knockdown. (f) *EVI1* expression levels relative to *PBGD* by qPCR upon *EVI1* knockdown in two different K562 *EVI1-eGFP* clones (clones 8 and 24) and WT K562. Statistical test: one-way ANOVA. The error bar represents the standard deviation (SD). (g) *eGFP* expression levels relative to *PBGD* by qPCR upon *EVI1* knockdown in two different K562 *EVI1-eGFP* clones (clones 8 and 24) and WT K562. Statistical test: one-way ANOVA. The error bar represents the standard deviation (SD). (h) Sorting strategy flow cytometry and FACS sort experiments. A control is shown as an example (clone 8-4, no CRISPR-Cas9 targeting). Gate selects the viable cells. (i) Similar to h. Gate selects the single cells. (j) Similar to h. Gate selects the eGFP negative and eGFP positive cells. (k) Similar to h. Histogram shows eGFP levels in the FITC-A channel. (l) Sorting strategy flow cytometry and FACS sort experiments. An experiment targeting the CTCF binding site near the *EVI1* promoter is shown as an example (clone 8-4 CRISPR-Cas9 targeted with CTCF sgRNA#1). Gate selects the viable cells. (m) Similar to l. Gate selects the single cells. (n) Similar to l. Gate selects the eGFP negative and eGFP positive cells. These gates were used for FACS sorting experiments. Figures 4c, 4d, 4e, 4g, 4h, 5c-f, 6c-g and Supplementary Figures S3a, S5a-c. (o) Similar to l. Histogram shows eGFP levels in the FITC-A channel. (p) An overlay of the eGFP levels of the control and CRISPR-Cas9 targeted experiments is shown displaying eGFP levels as histograms. This gating and display strategy was used for Figures 2b, 2d, 2g, 2h, 4b, 5b, 6b and Supplementary Figures S1e, S2g, S4a, S4g and S7a.



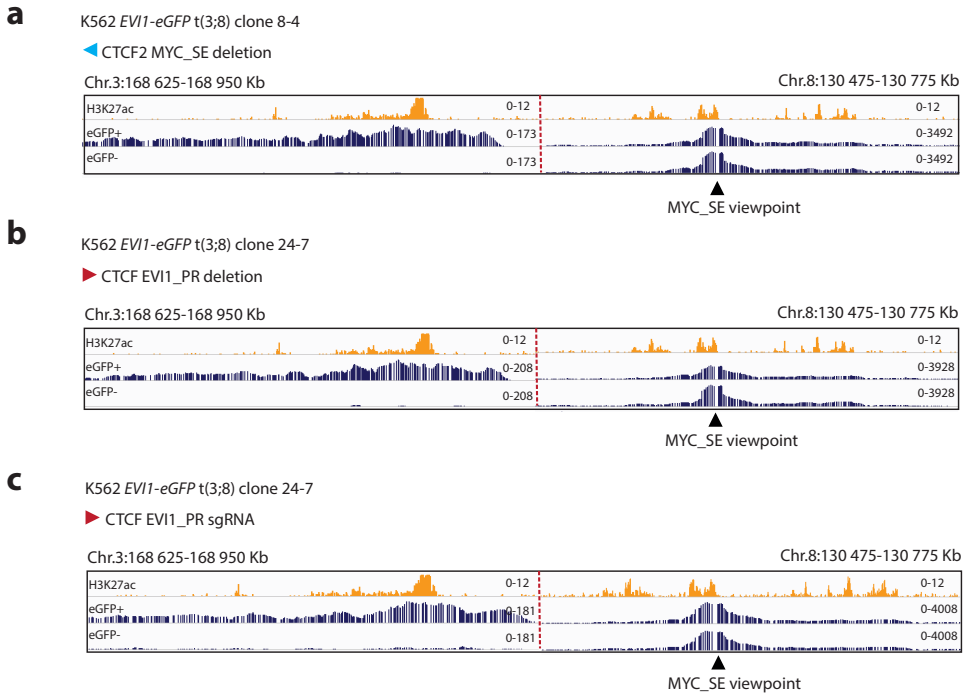
Supplementary Figure 2. A t(3;8) cell model recapitulates *EVI1* overexpression in human AML. (a) Sanger sequencing (amplicon covering the breakpoint as shown in Figure 2G), to validate the generation of a t(3;8). Nucleotide sequencing data of all four K562 *EVI1-eGFP* t(3;8) clones is shown. Using the online tool BLAT [1] the left part of the sequence maps back to 3q26.2 (pink) and the right part maps to 8q24 (blue). About 100bp were deleted on the Chr.8 side of the breakpoint in the generation of the translocation (depicted below in red). (b) Schematic overview of diagnostic FISH experiments performed to validate the presence of t(3;8) in the four clones. Fluorescent FISH pictures of K562 t(3;8) clone 8-3 are in C-E are shown. (c) Detection of t(3;8) by FISH. *MECOM* was split and the red signal was separated from the blue and green, indicating a translocation of the distal part of the q arm of chromosome 3. Scale bar 5 µm. (d) Fluorescent images obtained by FISH to detect t(3;8). Three red signals represent the *MYC* gene on chromosomes 8 (3 copies of Chr.8 in K562), demonstrating that the *MYC* gene is unaffected. The longer tip of the derivate Chr.8 near the *MYC* signal (arrow) is in line with a (t(3;8)) rearrangement at this chromosome. Scale bar 5 µm. (e) Fluorescent images obtained by FISH to detect t(3;8). The three Chr.8s can be visualized with the bigger red signal at the centromeres; the red part of the separated *MECOM* probe is located at the q-arm of one of the Chr.8s (arrow). Together the FISH images demonstrate that part of *MECOM* had been translocated to Chr.8, forming a t(3;8)(q26;q24). Scale bar 5 µm. (f) Western blot showing high EVI1 protein levels for t(3;8) clone 8-4, and a reduction of EVI1 levels upon *EVI1* knockdown (KD) using a shRNA. Source data are provided as a Source Data file. (g) Flow cytometry plot showing eGFP reduction upon *EVI1* knockdown.



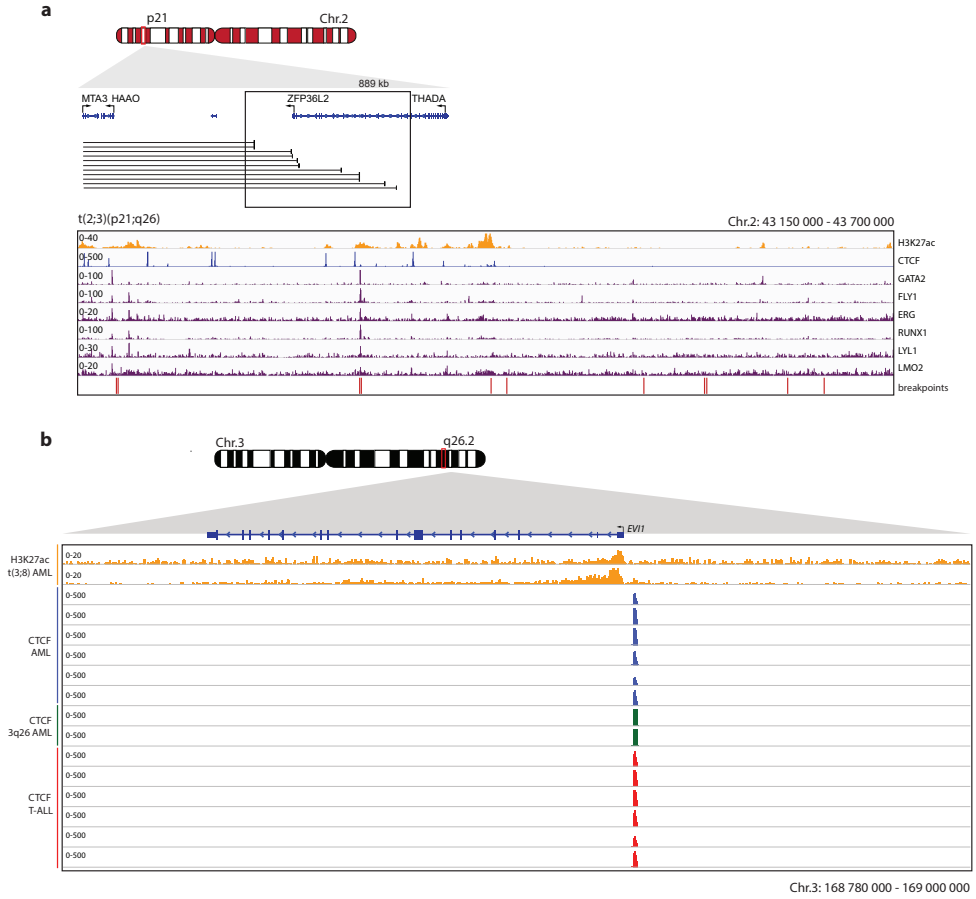
Supplementary Figure 3. One critical enhancer module in the *MYC* SE drives *EVI1* transcription. (a) PCR showing the deletion of enhancer module C present in the bulk cells, eGFP- and eGFP+ fractions, but not in the controls. The deletion observed by the smaller PCR band in the eGFP+ fraction can be explained by deletions in the non-translocated Chr.8 allele not influencing *EVI1* expression, and thus not eGFP. Nevertheless, clearly more cells with the deletion are found in the eGFP- fraction. Source data are provided as a Source Data file. (b) Boxplots depicting the accessibility and activity of the *MYC* SE modules as measured by ATAC-seq (N=5) and H3K27ac ChIP-seq (N=4) in t(3;8) AML patients. The Y axis indicates the Reads Per Kilobase of transcript per Million mapped reads (RPKM) for each module. Statistically significant differences were determined by a Kruskal-Wallis test. Both ATAC-seq and H3K27ac showed that module C is distinctly active. The lower and upper edges of the boxplots represent the first and third quartiles, respectively, the horizontal line inside the box indicates the median. The whiskers extend to the most extreme values within the range comprised between the median and 1.5 times the interquartile range. The circles represent outliers outside this range. (c) *MYC* SE module C is active and bound by HSPC-active transcription factors in t(3;8) AML (orange tracks), t(3;8) K562 (dark blue), wild type K562 (light blue) and CD34+ cells (red). For each of these groups, ATAC-seq, H3K27ac ChIP-seq and transcription factor ChIP-seq are shown. Transcription factor ChIP-seq data for CD34+ and wild type K562 cells are publicly available from [2] and [3] respectively. (d) 4C-seq data of t(3;8) clone 24-7 (blue tracks) with the *EVI1* promoter as viewpoint (black triangle). No interaction changes were observed following the deletions of *MYC* SE modules D or I (Figure 4B). The top H3K27ac ChIP-seq track (orange) shows the presence of the active *EVI1* promoter and the modules of the *MYC* SE. (e) 4C-seq data of t(3;8) clone 24-7 (blue tracks) with viewpoint in the *MYC* SE (black triangle) showing the same as D. No interaction changes upon deletion of *MYC* SE modules D or I were observed (Figure 4B).



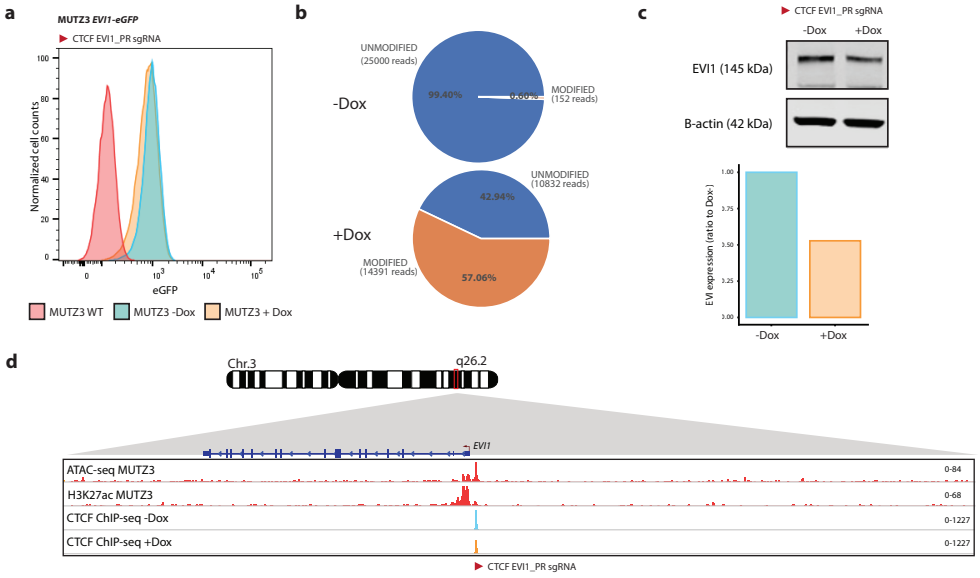
Supplementary Figure 4. *EVI1* is not regulated by *MYC* SE in a control model with 3q/*MECOM* amplification. (a) Flow cytometry plot comparing eGFP levels of *EVI1-eGFP* K562 with 3q/*MECOM* amplification (*EVI1*-AMP) clone 24-2 to WT, parental (*EVI1-eGFP*) and t(3;8) clone 24-7 K562 cells. (b) Western blot comparing EVI1 protein levels of *EVI1*-AMP clones 24-1 and 24-2 to WT, parental (*EVI1-eGFP*) and t(3;8) clones 24-7 and 24-12. Source data are provided as a Source Data file. (c) *MECOM (EVI1)* FISH for clone 24-2 illustrating the *EVI1* amplification. Scale bar 5 µm. (d) Overview of Chr.3 SNP array data showing high copy number for the q arm of Chr.3 starting from 3q26.2 in clone 24-2. Copy number gains are indicated in blue, whereas copy number losses are indicated in red. (e) SNP array data: zoom-in on the 3q26.2 locus of clone 24-2 indicating the breakpoint/start of the amplification (blue area). The amplification includes exactly the complete *EVI1 (MECOM)* locus. We estimated clone 24-2 has 4 copies of this part of the 3q arm including *EVI1*, resulting in elevated *EVI1* expression and protein levels. (f) Clone 24-2 was used as a control clone with high *EVI1* expression but without a t(3;8). All genomic deletions in the *MYC* SE (Figure 4A and 5A) made in this study were also successfully performed in clone 24-2 as shown here by PCR. Flow analysis showed that none of the deletions produced changes in *EVI1* expression in this clone.



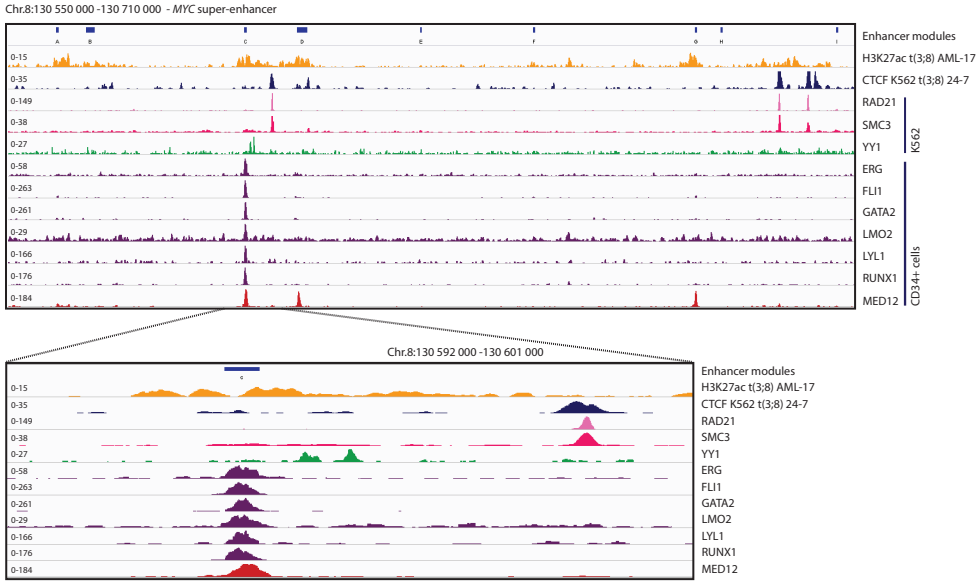
Supplementary Figure 5. CTCF binding site upstream of the *EVI1* promoter hijacks a *MYC* SE in t(3;8) AML. (a) 4C-seq data (t(3;8) clone 8-4), with the *MYC* SE as viewpoint, of cells with a CTCF2 *MYC* SE deletion and sorted on eGFP. In the eGFP- fraction (lower blue), a loss of interaction with the *EVI1* promoter was observed compared to the eGFP+ cells (upper blue). The top H3K27ac ChIP-seq track (orange) shows the presence of the active *EVI1* promoter and the modules of the *MYC* SE. (b) 4C-seq data (t(3;8) clone 24-7), with the *MYC* SE as viewpoint, of cells with a CTCF *EVI1* promoter deletion and sorted on eGFP. In the eGFP- fraction (lower blue), a loss of interaction with the *EVI1* promoter was observed compared to the eGFP+ cells (upper blue). The top H3K27ac ChIP-seq track (orange) shows the presence of the active *EVI1* promoter and the modules of the *MYC* SE. (c) Comparison of chromatin interaction at the *EVI1* promoter for eGFP+ (upper blue) and eGFP- (lower blue) cells, shown by 4C-seq (t(3;8) clone 24-7) with the *MYC* SE as viewpoint after targeting the CTCF motif with the sgRNA as presented in Figure 6D.



Supplementary Figure 6. Translocated region in t(2;3)(p21;q26) AML contains strong regulatory elements. (a) Translocated locus in t(2;3)(p21;q26) AML showing the presence of the gene *THADA*, the black box indicating the zoom-in shown below. Zoom-in: ChIP-seq for H3K27ac (t(3;8) patient AML-17, orange) indicating putative enhancer regions, CTCF binding and HSPC-active transcription factor recruitment. The red lines below indicate the exact breakpoint of the t(2;3) AMLs (detected by 3q-seq). We predict that the translocated region of each of these cases could be unique, but in all cases contains a strong regulatory locus. (b) ChIP-seq data showing H3K27ac in 2 t(3;8) AML patients (yellow), CTCF binding in 6 non-3q26 AML cases (blue), 2 AML cases with 3q26 rearrangements (green) and in 6 T-ALL cases (red). *EVI1* is only expressed in the 3q26-rearranged AML cases (TPM > 1), comprising one t(3;8) and one inv(3) patient. Thus, CTCF binding seems to be constitutive and independent of *EVI1* expression.



Supplementary Figure 7. The CTCF binding site upstream of *EVI1* is also critical for enhancer hijacking in other models of 3q26-rearranged AML. (a) Flow cytometry plot showing the effect of targeting the CTCF *EVI1*_PR site in MUTZ3 cells with sgRNA. Cas9 was induced by doxycycline (Dox), leading to loss of eGFP (orange) compared to control cells (blue). In red, wild type MUTZ3 cells without the eGFP reporter. (b) Amplicon-seq data showing the percentage of modified (orange) and unmodified (blue) reads in MUTZ3 after targeting the *EVI1*_PR CTCF with sgRNA. Cas9 was induced by doxycycline (+Dox), leading to successful genome editing compared to control cells (-Dox). (c) Western blot showing depletion of EVI1 protein in MUTZ3 after targeting the CTCF *EVI1*_PR binding site with sgRNA. Source data are provided as a Source Data file. The barplot below presents the expression levels of EVI1 measured as EVI1/B-actin ratio, relative to expression in Dox- cells. The bars only represent one data point. (d) Epigenomic landscape of the *EVI1* promoter in MUTZ3 following the deletion of the CTCF *EVI1*_PR site. In red, ATAC-seq and H3K27ac ChIP-seq from wild type MUTZ3 cells indicating the presence of an active promoter. CTCF ChIP-seq shows a moderate loss of CTCF upon targeting of the CTCF *EVI1*_PR site with sgRNA and Cas9 induction (orange), compared to control cells without Cas9 induction (blue).



Supplementary Figure 8. Transcription factor and transcriptional co-factor occupation of the *MYC* SE. In the upper panel, the *MYC* SE and its distinct enhancer modules are illustrated; in the lower panel, a zoom-in on enhancer module C is shown. Active enhancer regions are illustrated by H3K27ac of the t(3;8) AML-17 patient (orange) and CTCF binding locations are marked by CTCF ChIP-seq data of t(3;8) K562 clone 24-7. ChIP-seq data in K562 WT of cohesin subunits RAD21 (light pink) and SMC2 (pink), as well as YY1 (green), were retrieved from ENCODE [4]. ChIP-seq data in CD34+ from healthy donors were retrieved from other publications, including a heptad of hematopoietic transcription factors [3] (ERG, FLI1, GATA2, LMO2, LYL1, and RUNX1) in purple and MED12 [5] in red.

SUPPLEMENTARY REFERENCES

1 Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002).

2 Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* **46**, D794-D801 (2018).

3 Beck, D. *et al.* Genome-wide analysis of transcriptional regulators in human HSPCs reveals a densely interconnected network of coding and noncoding genes. *Blood* **122**, e12-22 (2013).

4 Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

5 Aranda-Orgilles, B. *et al.* MED12 Regulates HSC-Specific Enhancers Independently of Mediator Kinase Activity to Control Hematopoiesis. *Cell Stem Cell* **19**, 784-799 (2016).

6 Groschel, S. *et al.* A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* **157**, 369-381 (2014).

7 Fernandez, J. M. *et al.* The BLUEPRINT Data Analysis Portal. *Cell Syst* **3**, 491-495 e495, doi:10.1016/j.cels.2016.10.021 (2016).

8 Smeenk, L. *et al.* Selective requirement of MYB for oncogenic hyperactivation of a translocated enhancer in leukemia. *Cancer Discov*, doi:10.1158/2159-8290.CD-20-1793 (2021).

9 Langmead, B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics* **Chapter 11**, Unit 11 17, doi:10.1002/0471250953.bi1107s32 (2010).

10 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

11 Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165, doi:10.1093/nar/gkw257 (2016).

12 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137-R137, doi:10.1186/gb-2008-9-9-r137 (2008).

13 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).

14 Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419, doi:10.1038/nmeth.4197 (2017).

15 Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521, doi:10.12688/f1000research.7563.2 (2015).

16 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

17 Trim Galore (2012).

18 Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner. (2014).

19 Clement, K. *et al.* CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat Biotechnol* **37**, 224-226, doi:10.1038/s41587-019-0032-3 10.1038/s41587-019-0032-3 [pii] (2019).

20 Ziebarth, J. D., Bhattacharya, A. & Cui, Y. CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization. *Nucleic Acids Res* **41**, D188-194, doi:10.1093/nar/gks1165 (2013).

21 Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **48**, D87-D92, doi:10.1093/nar/gkz1001 (2020).

CHAPTER 5

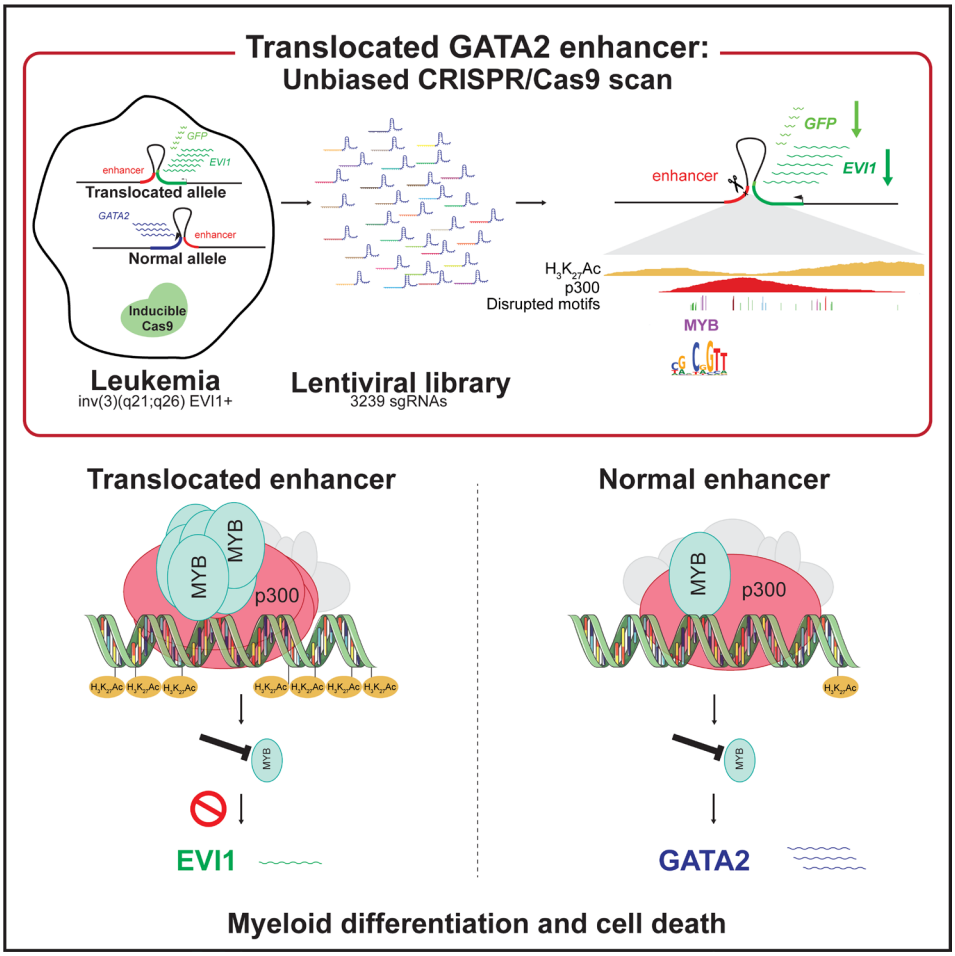
Selective requirement of MYB for oncogenic hyperactivation of a translocated enhancer in leukemia

Leonie Smeenk[1,2], Sophie Ottema[1,2], **Roger Mulet-Lazaro[1,2]**, Anja Ebert[5], Marije Havermans[1,2], Andrea Arricibita Varea[1,2], Michaela Fellner[5], Dorien Pastoors[1,2], Stanley van Herk[1,2], Claudia Erpelinck-Verschueren [1,2], Tim Grob[1], Remco M. Hoogenboezem[1], François G. Kavelaars[1], Daniel R. Matson[3], Emery H. Bresnick[3], Eric M. Bindels[1], Alex Kentsis[4], Johannes Zuber[5,6,7] and Ruud Delwel[1,2,7]

[1] Department of Hematology, Erasmus MC Cancer Institute, Rotterdam, the Netherlands
[2] Oncode Institute, Utrecht, the Netherlands
[3] Department of Cell and Regenerative Biology, University of Wisconsin – Madison, Madison, WI, USA
[4] Tow Center for Developmental Oncology, Sloan Kettering Institute, Department of Pediatrics, Weill Medical College of Cornell University, Memorial Sloan Kettering Cancer Center, New York, USA
[5] Research Institute of Molecular Pathology (IMP), Vienna BioCenter (VBC), Vienna, Austria
[6] Medical University of Vienna, Vienna BioCenter (VBC), Vienna, Austria
[7] These authors contributed equally

Running title: MYB is required for a hijacked enhancer in AML

Cancer Discov. 2021 Nov;11(11):2868-2883. doi: 10.1158/2159-8290.CD-20-1793.
PMID: 33980539


Translocated GATA2 enhancer:
Unbiased CRISPR/Cas9 scan
GFP
EVI1
enhancer
Translocated allele
GATA2
enhancer
Normal allele
Inducible Cas9
Leukemia
inv(3)(q21;q26) EVI1+
Lentiviral library
3239 sgRNAs
GFP
EVI1
enhancer
H3K27Ac
p300
Disrupted motifs
MYB
Translocated enhancer
MYB
MYB
p300
EVI1
Normal enhancer
MYB
p300
GATA2
Myeloid differentiation and cell death

ABSTRACT

In acute myeloid leukemia (AML) with inv(3)(q21;q26) or t(3;3)(q21;q26), a translocated *GATA2* enhancer drives oncogenic expression of *EVI1*. We generated an EVI1-GFP AML model and applied an unbiased CRISPR/Cas9 enhancer scan to uncover sequence motifs essential for *EVI1* transcription. Using this approach, we pinpointed a single regulatory element in the translocated *GATA2* enhancer that is critically required for aberrant *EVI1* expression. This element contained a DNA binding motif for the transcription factor MYB which specifically occupied this site at the translocated allele and was dispensable for *GATA2* expression. *MYB* knockout as well as peptidomimetic blockade of CBP/p300-dependent MYB functions resulted in downregulation of EVI1 but not of GATA2. Targeting *MYB* or mutating its DNA-binding motif within the GATA2 enhancer resulted in myeloid differentiation and cell death, suggesting that interference with MYB-driven *EVI1* transcription provides a potential entry point for therapy of inv(3)/t(3;3) AMLs.

STATEMENT OF SIGNIFICANCE

We show a novel paradigm in which chromosomal aberrations reveal critical regulatory elements that are non-functional at their endogenous locus. This knowledge provides a rationale to develop new compounds to selectively interfere with oncogenic enhancer activity.

INTRODUCTION

Next-generation sequencing has greatly improved our knowledge about the location, distribution and frequency of recurrent gene mutations in cancer [1,2]. The focus has previously been on the identification and understanding of mutations in protein coding regions. However, many mutations are found in intergenic regions as well [3], which now receive broad attention [4-8]. Those studies demonstrate that malignant transformation does not only rely on coding mutations in proto-oncogenes, but may also depend on aberrant regulation of oncogene expression. Well-described mechanisms of aberrant gene activation include generation of novel enhancers by nucleotide substitution, focal amplification of enhancers, loss of boundaries between topologically associated domains (TAD) or enhancer hijacking by chromosomal rearrangements [9-15].

Chromosomal inversion or translocation between 3q21 and 3q26 (inv(3)(q21;q26) or t(3;3)(q21;q26)) in AML result in the aberrant expression of the proto-oncogene *EVI1* located at the *MDS1* and *EVI1* complex locus (*MECOM*) at 3q26 [16-19]. Our group and others reported that hyper-activation of *EVI1* is caused by a *GATA2* enhancer translocated from chromosome 3q21 to *EVI1* on chromosome 3q26 [12,20]. Upon translocation, this hijacked *GATA2* enhancer appears to behave as a super-enhancer and is marked by a broad stretch of H_3K_{27} acetylation [12,20]. In the current study, we aimed to unravel the mechanism by which the hijacked *GATA2* super-enhancer leads to *EVI1* activation. We generated a model to study *EVI1* regulation in inv(3)/t(3;3) AML cells by inserting a *GFP* reporter 3' of endogenous *EVI1* and introduced an inducible Cas9 construct. To uncover important elements in this hijacked enhancer, we applied CRISPR/Cas9 scanning and identified motifs essential for driving *EVI1* transcription. We demonstrated a single regulatory element in the translocated GATA2 enhancer that is critical for the regulation of *EVI1* expression, with an essential role for MYB through binding to the translocated enhancer. Treatment of inv(3)/t(3;3) AML cells with peptidomimetic MYB:CBP/p300 inhibitor decreased EVI1 expression, and induced leukemia cell differentiation and cell death.

RESULTS

Expression of *EVI1* in inv(3)/t(3;3) AML is reversible

In inv(3)/t(3;3) AMLs the *GATA2* super-enhancer is translocated to *MECOM*, driving expression of *EVI1* [12,20]. We investigated whether *GATA2* enhancer-driven transcription of *EVI1* in inv(3)/t(3;3) is reversible in leukemia cells. In primary inv(3)/t(3;3) AML, immature $CD34^+CD15^-$ cells can be discriminated from more mature $CD34^-CD15^-$ and $CD34^-CD15^+$ cells (Figure 1A, left and Figure S1A,B, left). Whereas EVI1 is highly expressed in $CD34^+CD15^-$ cells, mRNA and protein levels decline in the $CD34^-CD15^-$ fraction and are almost completely lost in $CD34^-CD15^+$ cells in inv(3)/t(3;3) primary AML as well as in MUTZ3 cells, an inv(3) AML model (Figure 1A,B, right, Figure 1C, Figure S1A,B, right). Since 3q26 rearrangements are present in all fractions, as determined by three-colored Fluorescent in-situ hybridization (FISH) (Figure S1C), we conclude that transcription of *EVI1* can be reversed in AML cells despite the presence of a 3q26 rearrangement. *In vitro* culture of sorted MUTZ3 cells revealed that only the *EVI1*-expressing $CD34^+CD15^-$ cells were competent to proliferate (Figure S1D), in agreement with previous observations showing that *EVI1* depletion results in loss of colony formation and induction of differentiation [12]. Thus, although AML cells with inv(3)/t(3;3) depend on *EVI1*, transcription of this gene remains subject to regulation and can be repressed, with major consequences for cell proliferation and differentiation.

Generation of an EVI1-GFP inv(3) AML model

Our findings indicate that interference with *EVI1* transcription may be an entry point to specifically target inv(3)/t(3;3) AMLs. To study the molecular mechanisms of *EVI1* transcriptional activation by the hijacked *GATA2* enhancer, we introduced a *GFP* reporter 3' of *EVI1* at the translocated allele, which is the only allele expressed in MUTZ3 cells. A T2A self-cleavage site was introduced in between *EVI1* and *GFP* separating the two proteins (Figure 2A, Figures S2A and S2B). Knockdown of *EVI1* using two unique *EVI1*-specific shRNAs (Figure 2B) resulted in a reduction of the GFP signal (Figure 2C). Subsequently, a construct with tight doxycycline (Dox) controlled expression of Cas9 was introduced into MUTZ3-EVI1-GFP cells (Figure S2C-D) and used to target the translocated *GATA2* enhancer and study *EVI1* regulation. Deletion of approximately 1000 bp in the -110 kb (-77 kb in mouse) **distal** *GATA2* enhancer [21,22] using two specific sgRNAs (Table S1) resulted in a severe decrease in *GFP* expression upon Dox treatment (Figure 2D). We sorted the GFP expressing cells into three fractions and observed that enhancer deletion was most pronounced in the GFP^{low} FACS-sorted cells (Figure 2E lower band). The GFP^{low} fraction also contained the lowest *GFP* and *EVI1* mRNA levels (Figure 2F-G). Cells from the GFP^{low} fraction, which showed reduced EVI1 expression, formed less colonies than GFP^{high} cells in methylcellulose (Figure 2H). Only colonies obtained from the GFP^{high} fraction consisted of cells able to multiply when placed in liquid culture (Figure S2E). Immunophenotyping of the colonies revealed that GFP^{high} fractions predominantly consisted of immature $CD34^+CD15^-$ cells, while in contrast

the GFP^{low} fraction contained the highest number of differentiated CD15^{+}CD34^{-} cells (Figure S2F). Together, this established a Dox-inducible Cas9 expressing inv(3)/t(3;3) AML model (MUTZ3-EVI1-GFP) for studying the transcriptional control of *EVI1* via a GFP reporter.

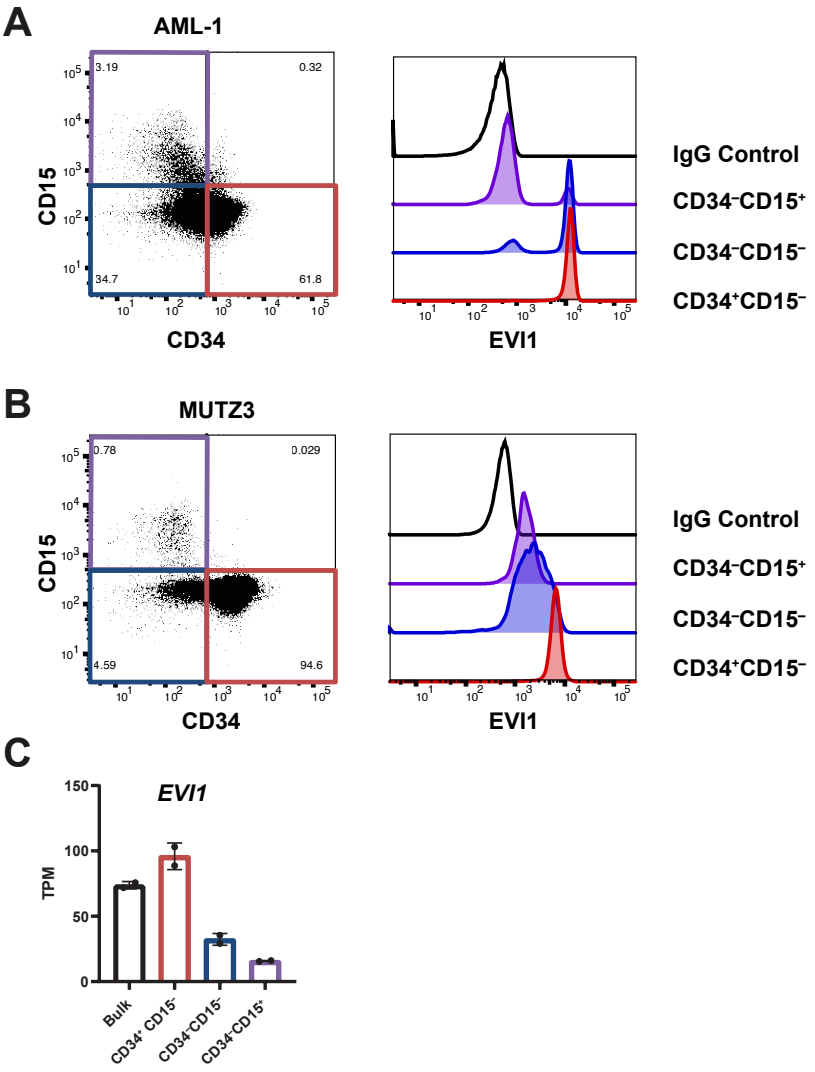


Figure 1. Expression of *EVI1* in inv(3)/t(3;3) AML is reversible. (A) Flow cytometric analysis of CD34- and CD15-stained inv(3;3) primary AML cells (AML-1) (left) and intracellular EVI1 staining in the gated fractions (right). (B) Flow cytometric analysis of MUTZ3 cells stained with CD34 and CD15 (left) and intracellular EVI1 staining in the gated fractions (right). (C) Bar plot showing relative expression of *EVI1* in Transcripts Per Million (TPM) in sorted fractions of MUTZ3 cells. Error bars represent standard deviation of two biological replicates.

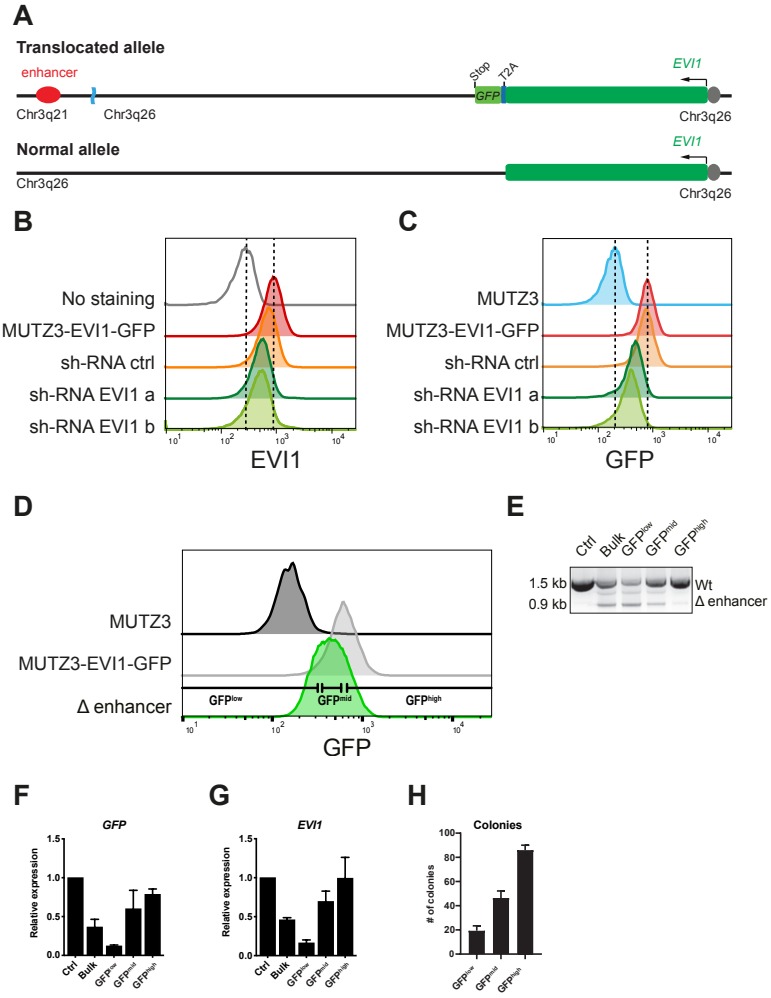


Figure 2. Generation of an EVI1-GFP inv(3) AML model. (A) Schematic representation of EVI1-GFP knock-in with a T2A self-cleavage site in the MUTZ3 cells at the endogenous translocated *EVI1* locus. (B) Flow cytometric analysis of intracellular EVI1 after shRNA-mediated knockdown of *EVI1* using two different shRNAs. The effects on EVI1 protein were measured 48 hours after transduction. Scrambled shRNAs were used as control. (C) Flow cytometric analysis of GFP in the same experiment indicated in (B). (D) Representative flow cytometric plot showing the effect of the -110kb *GATA2* enhancer deletion in MUTZ3-EVI1-GFP cells (Δ enhancer). Cas9 was induced with Dox 24h before nucleofection of two sgRNAs. The effect on EVI1 was measured by GFP levels using flow cytometric analyses. Cells were sorted 48h after nucleofection of subsequent sgRNAs into three fractions: GFP^{low}, GFP^{mid} and GFP^{high}. (E) Genotyping PCR showing a wild type (WT) band (1500 bp) or a band for the enhancer deleted(Δ) (900 bp), either in bulk (before sorting) or in sorted fractions. Control (Ctrl) represents PCR after nucleofection of the sgRNAs without Dox induction. (F) Bar plot showing relative *GFP* expression of bulk and sorted fractions analyzed by qPCR. The expression levels of *PBGD*, a housekeeping gene, were used as control for normalization. Relative expression is calculated as fold over Ctrl (nucleofection of the sgRNAs without Dox). Error bars represent standard deviation of two biological replicates. (G) Bar plot showing relative *EVI1* expression of MUTZ3-EVI1-GFP bulk and sorted fractions analyzed by qPCR. For details see Figure 2F legend. (H) Bar plot showing the number of colonies grown in methylcellulose from each sorted fraction. Colonies were counted 1.5 weeks after plating. Error bars represent standard deviation of three plates.

Unbiased CRISPR/Cas9 enhancer scan reveals a specific 1 kb region as essential for *EVI1* activation

The minimally translocated region of the *GATA2* super-enhancer is 18 kb long [12]. In MUTZ3 and MOLM1, which are both inv(3) AML models, this highly H_3K_{27} acetylated region (Figure 3A; yellow) contains four loci of open chromatin determined by ATAC-seq (Figure 3A; orange), of which two show strong p300 occupancy (Figure 3A; red). To identify, in an unbiased fashion, which elements of the 18 kb translocated region control *EVI1* transcription, we employed a CRISPR/Cas9-based enhancer scanning approach (Figure 3A). We constructed a lentiviral library containing 3239 sgRNAs covering the 18 kb translocated region (Figure 3A, Table S2) and transduced it into MUTZ3-EVI1-GFP cells at a low multiplicity of infection. After neomycin selection and cell expansion, the cells were treated with Dox to induce Cas9 expression and cells displaying reduced GFP reporter expression (GFP^{low}) were selected by flow cytometric sorting at day 5 and day 7. The sgRNAs were amplified from genomic DNA and deep-sequenced to identify the sgRNAs that were enriched in the GFP^{low} fraction. The log2fold change of 3 independent experiments were combined as shown in Figure 3B, which demonstrated a strong correlation between the sgRNAs enriched in GFP^{low} cells at day 5 and day 7 (Figure 3B).

Five sgRNAs targeting *EVI1* were the top scoring hits in the GFP^{low} fraction (indicated in blue), whereas sgRNAs targeting the safe harbor AASV1 locus (in red) were not enriched, emphasizing the specificity and sensitivity of the assay (Figure S3A). sgRNAs with a minimum of 3-fold enrichment in the GFP^{low} fraction all clustered in a small region of approximately 700 bp (Figure 3C). This region is a known p300-interacting region, which belongs to the -110 kb **distal** *GATA2* enhancer [21,22]. This p300-interacting region is occupied by a heptad of transcription factors (SCL, LYL1, LMO2, GATA2, RUNX1, FLI1 and ERG) that regulate gene expression in hematopoietic stem and progenitor cells (HSPCs) [23,24] (Figure S3B). Approximately 40 sgRNAs within this region, with at least a 2-fold enrichment in the GFP^{low} fraction, were selected and cloned into a lentiviral construct with iRFP720 for individual testing. The loss of GFP signal at day 7 in the $iRFP^{+}$ fraction (gating strategy, see figure S3C) highly correlated with the enrichment of those 40 sgRNAs in the GFP^{low} fraction as observed in the enhancer scan (Figure 3D). An efficiently cutting sgRNA that was not enriched in the enhancer scan did not affect GFP signal upon Dox exposure (Figure S3D,E). Deep amplicon sequencing of the -110 kb enhancer region upon targeting by 36 individual sgRNAs revealed frequent mutations in motifs for MYB, GATA, RUNX-, MEIS-, XBP- and ETS- binding sites, which were among the highest conserved (Figure 3E, Figure S3F, Table S3).

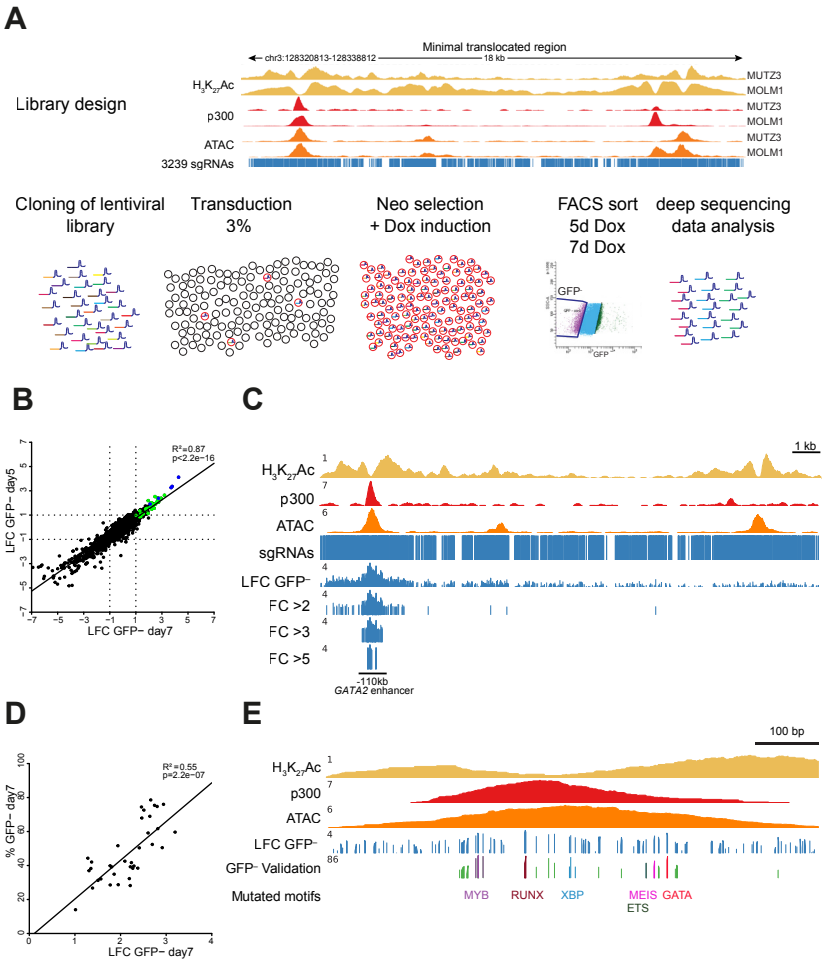


Figure 3. Unbiased CRISPR/Cas9 enhancer scan reveals one 1 kb region to be essential for *EVI1* activation. (A) ChIP-seq to determine $H_3K_{27}Ac$ pattern and p300 binding as well as open chromatin analysis using ATAC-seq in MUTZ3 and MOLM1 cells. The locations of the >3200 sgRNAs targeting the enhancer are indicated as vertical blue lines. A schematic overview of the enhancer scanning strategy is depicted below. (B) Scatter plot of enrichment of sgRNAs in sorted GFP^{low} fractions at day 5 and day 7 upon Dox induction. The average of three independent experiments for each dot is depicted. For every sgRNA detected in the GFP^{low} fractions the log2fold change (LFC) of the +Dox relative to –Dox was calculated. Five sgRNAs targeting *EVI1* were added to the sgRNA library as positive controls and are indicated in blue. The sgRNAs selected for further validation are indicated in green. The fitted linear regression and corresponding R-squared and p-value are indicated. (C) The LFC enrichment at day 7 of all sgRNAs and of sgRNAs with >2, >3 or >5 fold enrichment of sgRNAs in the GFP^{low} fractions at the 18 kb region of the *GATA2* super-enhancer in MUTZ3 cells is depicted. The $H_3K_{27}Ac$ pattern, p300 binding, open chromatin (ATAC) and location of all sgRNAs are indicated to visualize which sgRNAs were enriched in the GFP^{low} fraction. The -110 kb **distal** *GATA2* enhancer is indicated. (D) Scatter plot showing enrichment of sgRNAs in sorted GFP^{low} fractions at day 7 compared to $\%GFP^{neg}$ cells at day 7 for individually validated sgRNAs (based on two independent biological experiments). The sgRNAs used for validation are indicated by dots. The fitted linear regression and corresponding R-squared and p-value are indicated. (E) Zoom-in of the -110 kb *GATA2* enhancer (chr3:128322411-128323124) showing $H_3K_{27}Ac$ pattern, p300 binding and open chromatin (ATAC), LFC enrichment of sgRNAs at day 7 and the $\%GFP^{neg}$ cells at day 7 of the individually validated sgRNAs. Mutations in motifs for known transcription factors identified in the individually validated sgRNAs are indicated.

A MYB binding motif is essential for *EVI1* rather than for *GATA2* transcription

Four sgRNAs, i.e. # 3, 8, 11 and 16, generating the highest GFP^{neg} ($EVI1^{neg}$) fraction in the single guide validation experiments, all targeted the same region containing a potential MYB-binding motif (Figure 4A). The strong reduction of GFP expression, as tested for three of those guides (Figure 4B), was accompanied by loss of EVI1 protein (Figure 4C) and mRNA (Figure 4D). EVI1 loss was accompanied by differentiation into $CD34^{-}CD15^{+}$ cells in the sgRNA8-targeted GFP^{low} fraction (Figure 4E), in line with the findings in primary AML cells (Figure 1A,B, left and Figure S1A,B, left). Strikingly, sgRNA8-directed mutations within the enhancer did not affect GATA2 protein (Figure 4C) or mRNA levels (Figure 4D). Western blot analysis on sorted fractions of sgRNA8-treated cells revealed a strong reduction of EVI1 but not of GATA2 in GFP^{low} cells (Figure 4F). Amplicon-seq within the GFP^{low} sorted fraction of sgRNA8-treated cells revealed that almost 97% of the aligned sequences, including the translocated and non-translocated allele, were mutated (Figure 4G). In approximately 86% of all aligned sequences, the MYB motif was mutated. In 14%, a 20 bp deletion fully eliminated the predicted MYB DNA-binding motif (Figure 4H). We carried out pulldown experiments in which equal amounts of MUTZ3 nuclear lysates (Figure S4A) were exposed to beads with immobilized 100 bp enhancer DNA fragments representing WT or MYB-motif mutant enhancer DNA, as defined in Figure 4H. Western Blot analysis confirmed MYB binding to the 100 bp WT enhancer fragment (Figure 4I). MYB binding to the M1 or M2 mutants was severely reduced, but it was preserved in the M3 mutant, in which the MYB DNA-binding motif was retained (Figure 4I). We conclude that in inv(3)/t(3;3) AML transcription of *EVI1* depends on the presence of a MYB DNA-binding motif in the translocated enhancer. Strikingly, this MYB motif appears less relevant for the transcription of *GATA2* in the non-translocated allele.

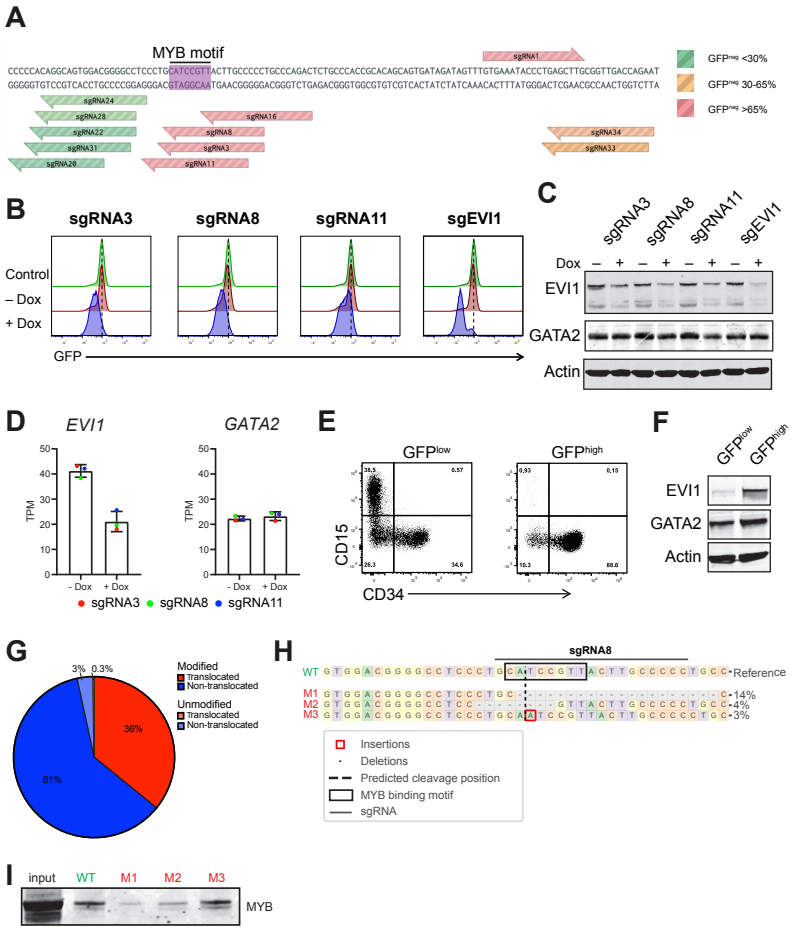


Figure 4. A MYB binding motif is essential for *EVI1* rather than for *GATA2* transcription. (A) Nucleotide sequence of the region targeted by sgRNAs 3,8,11 and 16, as well as other nearby sgRNAs, with the corresponding MYB DNA binding motif highlighted in purple. Colors of sgRNAs represent differences in percentage of recovery in the GFP^{neg} fraction. sgRNAs indicated in red are the most highly enriched in the GFP^{neg} fraction. (B) Flow cytometric analysis of MUTZ3-EVI1-GFP cells upon sgRNA treatment. GFP signal shifts are shown upon transduction with lentivirus containing sgRNAs 3, 8, 11 or an EVI1-specific sgRNA. Cells were analyzed by flow cytometry 7 days after induction of Cas9. (C) Western blot using EVI1- and GATA2-specific antibodies upon transduction with lentivirus containing sgRNAs 3, 8, 11 or an *EVI1* specific sgRNA (EVI1.4) analyzed 7 days after induction of Cas9. Actin was used as loading control. (D) Bar plot showing relative expression of *EVI1* and *GATA2* in transcripts per million (TPM) in MUTZ3-EVI-GFP cells treated with sgRNAs 3, 8 or 11, –Dox or +Dox. The cells treated with sgRNAs 3, 8, or 11 were considered replicates and standard deviation is shown. (E) CD34/CD15 flow cytometric analyses of MUTZ3 EVI1-GFP cells transduced with sgRNA8 (+Dox), sorted for GFP^{low} or GFP^{high} and analyzed two weeks after sorting. (F) EVI1 and GATA2 western blot upon treatment with sgRNA 8, sorted into GFP^{low} or GFP^{high} fractions, 7 days after induction of Cas9. Actin was used as loading control. (G) Editing frequency in the GFP^{low} fraction of sgRNA8-treated cells. Modified reads exhibited variations with respect to the reference human sequence. The percentages of reads that align to each allele were determined based on a heterozygous SNP in the sequenced region. (H) Visualization of the distribution of mutations identified around the sgRNA8 target site in the GFP^{low} sorted fraction. The sgRNA8 target site is indicated (GGGGGCAAGTAACGGATGC) as well as the MYB binding motif (black rectangle). (I) Western blot using anti-MYB antibody in MUTZ3 cell lysates following pulldowns using WT, mutated M1, M2 or M3 100bp DNA fragments.

Differential MYB binding and H_3K_{27} acetylation at the hijacked *GATA2* enhancer

ChIP-seq revealed MYB occupancy at the -110 kb *GATA2* enhancer in MUTZ3 and in inv(3)/t(3;3) AML patient cells (Figure 5A,B, green tracks). MYB also occupied the -110 kb *GATA2* enhancer in $CD34^+$ cells (Figure S4B, green track). Based on a heterozygous SNP in the -110 kb *GATA2* enhancer in MUTZ3, the translocated allele (*EVI1*) can be discriminated from the non-translocated (*GATA2*) allele [12]. We found approximately 7 times more MYB occupancy at the translocated allele (Figure 5A, right), in agreement with the finding that p300 occupancy (Figure 5A, red track) was also detected predominantly at the translocated enhancer (Figure 5A, right). Furthermore, H_3K_{27}Ac signal (Figure 5A, right) and open chromatin (ATAC) (Figure S4C) were 5 times more prevalent at the translocated enhancer. No SNPs were present in primary AMLs to discriminate MYB binding to the different alleles. However, based on two SNPs in the 18 kb region (Figure 5B, left), we observed a strong H_3K_{27}Ac allelic skewing of the primary inv(3)/t(3;3) AML, predicted to be biased to the translocated allele (Figure 5B, right). These data suggest that MYB and p300 interact with the -110 kb enhancer preferentially at the translocated allele. In sgRNA8-treated MUTZ3 cells (+Dox) MYB binding to the -110 kb site was significantly decreased compared to control (–Dox) cells (Figure 5C). This loss was *GATA2* enhancer-specific, since genome-wide MYB chromatin occupancy, which includes the MYB target gene *BCL2*, did not change in +Dox cells (Figure S4D and S4E). Importantly, the decrease of MYB-binding at the -110 kb enhancer upon sgRNA8 treatment was greater within the translocated allele (Figure 5C, right). Using Cut&Run we demonstrated that H_3K_{27}Ac was severely decreased at the enhancer in GFP^{low} sorted cells (Figure 5D, blue track) compared to GFP^{high} sorted cells (Figure 5D, green track) following sgRNA8 treatment. Moreover, SNP analysis revealed that the remaining H_3K_{27}Ac at the enhancer in GFP^{low} cells occurred predominantly at the non-translocated allele (*GATA2*) (Figure 5D, right). These data demonstrate that mutating the MYB binding motif at the translocated -110 kb enhancer decreases MYB binding, thus inactivating the enhancer and reducing *EVI1* transcription.

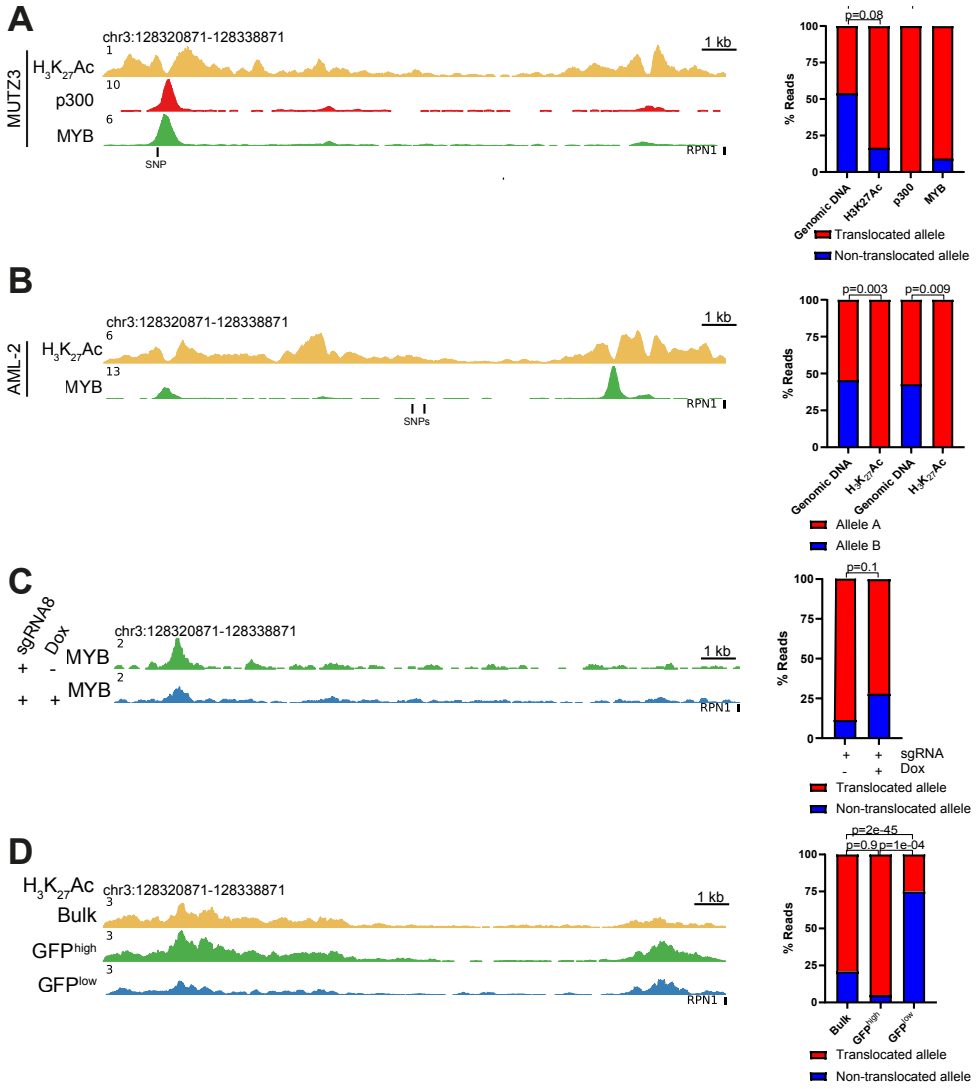


Figure 5. Differential MYB binding and H_3K_{27} acetylation at the hijacked *GATA2* enhancer. (A) H_3K_{27}Ac, p300 and MYB ChIP-seq profiles of the 18 kb super-enhancer region in MUTZ3 cells (left). Bar plot showing allelic bias towards the translocated allele for H_3K_{27}Ac, p300 and MYB occupancy by ChIP-seq analysis based on a SNP (rs553101013) (right). Previous sequencing showed that G represents the translocated allele and A the wild type allele [12]. P-values were calculated using a χ^2 test. (B) H_3K_{27}Ac and MYB ChIP-seq profiles of the 18 kb super-enhancer in an AML patient with inv(3) (AML-2) (left). Bar plot showing discrimination between H_3K_{27}Ac at the two *GATA2* enhancer alleles based on two SNPs (rs2253125 and rs2253144) (right). P-values were calculated using a χ^2 test. (C) MYB ChIP-seq profile of the 18 kb super-enhancer in sgRNA8-treated MUTZ3-EVI1-GFP cells plus or minus Dox treatment (left). Bar plot showing allelic distribution of MYB binding in sgRNA8 treated MUTZ3-EVI1-GFP cells plus or minus Dox treatment (right). P-values were calculated using a χ^2 test. (D) H_3K_{27}Ac profile of the 18 kb super-enhancer in sgRNA8-treated MUTZ3-EVI1-GFP cells, determined by Cut&Run in bulk, in GFP^{high} and in GFP^{low} sorted fractions (left). Bar plot showing allelic bias for H_3K_{27}Ac in the bulk, GFP^{high} and GFP^{low} fractions (right). P-values were calculated using a χ^2 test.

MYB interference downregulates EVI1 but not GATA2

MYB is expressed in MUTZ3 cells, regardless of their differentiation status (Figure S4F). To study whether MYB is important for *EVI1* expression, MYB-specific sgRNAs were introduced into MUTZ3-EVI1-GFP cells. At day 3 and 6 post-Dox induction, loss of MYB expression was evident, which was accompanied by a decrease of EVI1 protein (Figure 6A). In contrast, in line with the effects of mutating the MYB binding motif, knockout of MYB did not decrease GATA2 protein expression (Figure 6A). This suggests that MYB is not functioning upstream of GATA2 via this motif in inv(3) cells. When we either knocked out MYB or mutated the MYB DNA-binding motif with sgRNAs in K562 cells (Figure S4G), a model without a 3q26 rearrangement, we also did not see an effect on GATA2 protein levels (Figure S4H).

The activity of MYB can be repressed using the peptidomimetic inhibitor MYBMIM, which impairs the assembly of the MYB:CBP/p300 complex [25]. In MUTZ3 cells, treatment with 25 μM MYBMIM caused a 50% reduction of viable cells, whereas the inactive MYBMIM analog TG3 showed no effect (Figure S4I). Treatment of MUTZ3 cells with 20 μM MYBMIM strongly reduced EVI1 protein levels (Figure 6B) without impacting MYB levels (Figure 6B). Consistent with the MYB knockout experiment (Figure 6A), MYBMIM treatment did not alter GATA2 protein levels (Figure 6B). A two-day exposure of MUTZ3 cells to MYBMIM reduced the number of colonies in methylcellulose (Figure 6C). Flow cytometric analysis of MYBMIM-treated colony cells revealed increased maturation ($CD34^-CD15^+$ cells) in comparison with TG3-treated controls (Figure 6D). We next introduced a *FLAG-Evi1* retroviral construct [26] allowing for constitutive murine *Evi1* expression in MUTZ3 cells (Figure S4J). Loss of colony formation upon MYBMIM treatment was partly rescued by *Evi1* overexpression (Figure 6E). Similarly, the mild effect of MYBMIM on differentiation of MUTZ3 cells (Figure 6F, MYBMIM-EV) was reduced (Figure 6F, MYBMIM-Evi1). This indicates that the effect of MYB interference on MUTZ3 cells is at least partly mediated via *EVI1*. Moreover, whereas MYBMIM treatment did not reduce MYB protein, it decreased MYB occupancy at the *GATA2* enhancer (Figure 6G). p300 occupancy also decreased, but to a lesser extent than MYB (Figure 6G). MYB binding was reduced at several sites, including the *BCL2* enhancer (Figure S4K). MYBMIM, but not TG3, reduced viability of inv(3)/t(3;3) AML patient cells (n=3) (Figure 6H), and treatment of AML primary cells with MYBMIM reduced EVI1 protein levels without affecting levels of MYB or GATA2 (Figure 6I). Finally, MYBMIM affected neither GATA2 nor EVI1 levels in normal $CD34^+$ cells (Figure 6J), suggesting that MYB has no effect on the *GATA2* enhancer or on *EV1* in normal HSPCs. In contrast to MUTZ3 cells, MYBMIM did not reduce the number of $CD34^+$ colonies in methylcellulose (Figure S4L). Thus, targeting MYB represents a promising therapeutic possibility in the context of inv(3)/t(3;3) AMLs with *EVI1* overexpression.

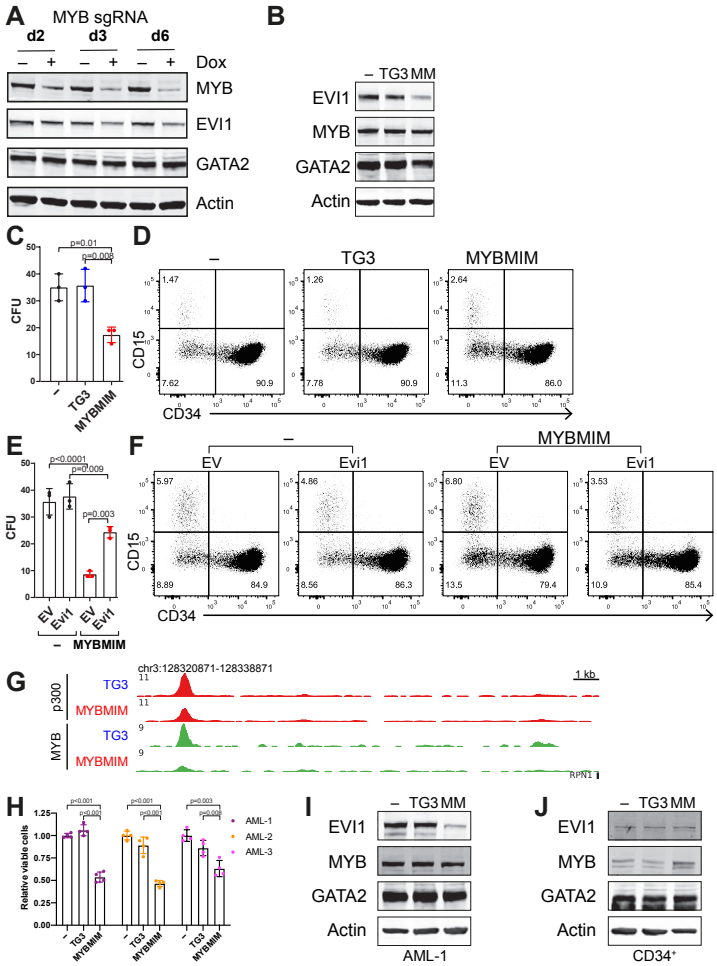


Figure 6. MYB interference downregulates EVI1 but not GATA2. (A) Western blot for MYB, EVI1 and GATA2 in MUTZ3-EVI1-GFP upon sgRNA-mediated MYB knockout (MYB.30) at indicated days after induction of Cas9. Actin was used as loading control. (B) Western blot for MYB, EVI1 and GATA2 in untreated cells (-) or cells treated for two days with 20 μM of TG3 or MYBMIM (MM). Actin was used as loading control. (C) Colony forming units (CFU) of MUTZ3 cells cultured without peptide or treated with 20 μM TG3 or MYBMIM for two days and subsequently plated in methylcellulose. Error bars show standard deviation across three plates. P-values were calculated using a one-way ANOVA test. (D) Flow cytometric analysis of MUTZ3 cells stained with CD34 and CD15. Cells studied by flow cytometry were either untreated or treated with 20 μM TG3 or MYBMIM for two days and subsequently grown for nine days in methylcellulose. (E) Colony forming units (CFU) of MUTZ3 cells with pMY-FLAG-Evi1-IRES-GFP (Evi1) or empty vector (EV) cultured without peptide or treated with 20 μM MYBMIM for two days and subsequently plated in methylcellulose. Error bars show standard deviation across three plates. P-values were calculated using a one-way ANOVA test. (F) Flow cytometric analysis of MUTZ3 cells with Evi1 or EV, stained with CD34 and CD15. Cells studied by flow cytometry were either untreated or treated with 20 μM MYBMIM for two days and subsequently grown for eight days in methylcellulose. (G) p300 and MYB ChIP-seq profiles of the 18 kb region in MUTZ3 cells treated with either 20 μM TG3 or MYBMIM for 48 h. (H) Cell-viability test of inv(3)/t(3;3) AML primary cells determined by CellTiter-Glo three days after culturing the cells in a 96-well plate with 20 μM TG3 or MYBMIM. Error bars show standard deviation across four biological replicates. P-values were calculated using a one-way ANOVA test. (I) Western blot for MYB, EVI1 and GATA2 in untreated AML cells or in AML cells treated with 20 μM TG3 or MYBMIM for 48h. Actin was used as loading control. (J) Western blot for MYB, EVI1 and GATA2 in cultured CD34$^+$ cells untreated or treated with 20 μM TG3 or MYBMIM for 48h. Actin was used as loading control.

DISCUSSION

Although multiple examples of hijacked enhancers causing uncontrolled expression of proto-oncogenes have been reported in various types of cancer [8,10,12,27,28], insight into their altered biological function remains limited. Elucidating these functions could provide opportunities for tailored interference and tools for therapeutic exploitation. Our unbiased CRISPR/Cas9 scan of the translocated 18 kb region in inv(3)/t(3;3) AMLs revealed a single region of approximately 1 kb essential for *EVI1* activation and leukemogenesis. This distal *GATA2* enhancer contained several conserved transcription factor DNA binding motifs, including an element preferentially occupied by MYB at the translocated allele. Strikingly, mutating this MYB binding motif in the enhancer at both alleles strongly decreased the expression of *EVI1,* but not of *GATA2*. GATA2 was also not affected in another leukemia line or in normal HSPCs. Together, these findings support a unique role for MYB in driving *EVI1* expression via the translocated enhancer, and suggest a potential vulnerability in inv(3)/t(3;3) AMLs. Indeed, peptidomimetic inhibition of MYB:CBP/p300 assembly in inv(3)/t(3;3) AML cells reduced EVI1 but not GATA2 protein levels, causing myeloid differentiation and cell death. This strengthens the hypothesis that interfering with *EVI1* expression via MYB may constitute a new entry point for targeting these AMLs. The fact that targeting MYB specifically compromises *EVI1* expression compared to *GATA2* points to the possibility of selectively targeting leukemia cells while sparing *GATA2* in normal HSPCs (Figure 6J), in which GATA2 is a vital regulator.

Although *MYB* encodes a transcription factor essential for normal hematopoiesis [29], there is also overwhelming evidence that it plays a critical role in malignant transformation. *MYB* was first discovered as an oncogene (v-*myb*) within the avian myeloblastosis virus (AMV) genome which generated myeloid leukemias in chickens [30,31]. Its critical involvement in super-enhancer activity was previously shown in human T-cell acute lymphoblastic leukemia (T-ALL) [9,32]. Mutations in non-coding regions near *TAL1* or *LMO2* create *de novo* binding sites for MYB, leading to the formation of new MYB-bound super-enhancers which drive uncontrolled transcription of those target genes. Furthermore, MYB binds to a translocated super-enhancer driving *MYB* expression in adenoid cystic carcinoma, creating a positive feedback loop sustaining its own expression [28]. MYB is also frequently overexpressed in human myeloid leukemias [33,34] and AML cells can be addicted to high levels of MYB and thus be more vulnerable to MYB inhibition than normal hematopoietic progenitor cells [35]. However, the mechanisms whereby MYB drives transformation to AML are not fully understood. To our knowledge, our results in this study represent the first example of a mechanism by which MYB drives oncogene activation in AML.

MYB occupies the translocated *GATA2* enhancer at a level considerably higher than the non-translocated enhancer. This may reflect increased chromatin accessibility as determined by $H_3K_{27}Ac$ ChIP-seq and ATAC-seq. The mechanisms driving this open chromatin pattern at the translocated locus remain a focus of future studies. However, translocation of the

enhancer to a new location places it in proximity to distinct promoters and regulatory elements which may ultimately impact chromatin accessibility and MYB binding. In support of this hypothesis, mutating the MYB DNA binding site or interference with MYB function causes reduced expression of EVI1 but not GATA2.

The coactivators CBP and p300 are major mediators of MYB transcriptional activity [36,37]. Therefore, specifically targeting the MYB:CBP/p300 interaction has been the focus of most small molecules seeking to inhibit MYB activity [25,38-41]. Experiments using the peptidomimetic inhibitor MYBMIM, which blocks the formation of MYB:CBP/p300 complex, showed a severe loss of EVI1 activity. As reported by Ramaswamy et. al., [25], we also observed that MYBMIM caused loss of MYB binding to the enhancer, with largely preserved total cellular levels of MYB. Concurrently, we observed that MYBMIM treatment did not inhibit p300 occupancy at the enhancer to the same extent as MYB occupancy. This partially retained p300 binding could be explained by the presence of other transcription factors bound at the *GATA2* enhancer that also recruit CBP/p300 (Figure S3B). MYBMIM reduced MYB binding at multiple sites which may be relevant in other leukemias in which MYB is essential [33-35]. Therefore, it is not surprising that other AML cell lines [25,42] respond to MYBMIM as well. While initial results with MYBMIM peptide treatment of inv(3)/t(3;3) AML cells are a promising proof of concept, MYBMIM peptide is very unstable *in vivo* (personal communication A.K.). Thus, development of small molecules with improved bioavailability that interfere with MYB:CBP/p300 complex will be required to investigate the relevance of MYB inhibition *in vivo*.

Our CRISPR/Cas9 scan identified one p300-interacting region containing a MYB DNA binding motif to be important for *EVI1* expression. Although mutations in the MYB DNA-binding motif had the biggest impact on *EVI1* expression, other mutations also reduced *EVI1* levels. These included mutations in consensus DNA binding sites for GATA-, RUNX-, MEIS-, XBP- and ETS-factors. Interestingly, some of these factors have been demonstrated to occupy the -110 kb enhancer in $CD34^+$ cells, including RUNX1, ERG and GATA2 [24]. MYB binding and activity at the -110 kb *GATA2* enhancer most likely occur in conjunction with p300 as well as transcription factors like RUNX1 and ERG. This is in accordance with other studies showing co-localization and potential cooperation between these factors and MYB [25,43,44]. Therefore, combinatorial targeting of MYB and other transcription factors may synergistically impact *EVI1* expression. This knowledge provides a rationale to develop new compounds to treat inv(3)/t(3;3) AML, which can be tested in our newly developed model.

Our findings provide important insight into the mechanisms of oncogenic enhancer-driven gene activation in AML. The selective MYB motif requirement for enhancer function at the translocated but not the normal allele constitutes a novel paradigm in which chromosomal aberrations reveal critical motifs that are non-functional at their endogenous locus. In principle, this paradigm may be extrapolated to other enhancer-driven cancers and even non-malignant pathologies.

METHODS

Data and Code Availability

Cell line sequence data generated in this study have been deposited at the EMBL-EBI ArrayExpress database (ArrayExpress, RRID:SCR_002964) under accession numbers E-MTAB-9939 (RNA-seq), E-MTAB-9949 (ATAC-seq), E-MTAB-9946 (Cut&Run-seq), E-MTAB-9945 (Amplicon-seq), E-MTAB-9948 (CRISPR enhancer scan) and E-MTAB-9959 (ChIP-seq). ChIP-seq and ATAC-seq data derived from donors or patients have been deposited at the European Genome-phenome Archive (The European Genome-phenome Archive (EGA), RRID:SCR_004944) under the accession number EGAS00001004839. This study did not generate any unique codes. All software tools used in this study are freely or commercially available.

Cell culture

The MUTZ3 cell lines (DSMZ Cat# ACC-295, RRID:CVCL_1433) were cultured in αMEM (HyClone) with 20% fetal calf serum (FCS) and 20% conditioned 5637 medium. The 293T (DSMZ Cat# ACC-635, RRID:CVCL_0063) were cultured in DMEM (Gibco) with 10% FCS. K562 (DSMZ Cat# ACC-10, RRID:CVCL_0004) was cultured in RPMI (Gibco) with 10% FCS. All cell lines were supplemented with 50 U/mL penicillin and 50 µg/mL streptomycin. Viable frozen AML cells and viable (frozen) bone marrow or cord blood $CD34^+$ cells were thawed and suspended in IMDM medium supplemented with: 20% BIT medium (StemCell Technologies), 1x β-mercaptoethanol (1000x Life technologies), 6 µg/ml LDL (Sigma Aldrich), human IL6, IL3, G-CSF, GM-CSF at 20 ng/ml and FLT3, SCF at 50 ng/ml (Peprotech). Cell lines were obtained from DSMZ and regularly confirmed to be mycoplasma-free by the MycoAlert Mycoplasma Detection Kit (Lonza, #LT07-318) according to the manufacturer's instructions.

Generation of model lines

The repair template was generated using Gibson Assembly (NEB). Both homology arms were PCR amplified from MUTZ3 genomic DNA using Q5 polymerase (NEB). The first homology arm consists of a part of the intron and last exon of *EVI1* minus the STOP codon. The second homology arm consists of part of the 3'UTR with the PAM sequence of sgRNA omitted. The T2A-eGFP was PCR amplified from dCAS9-VP64_2A_GFP (RRID:Addgene_61422). All fragments were cloned using Gibson assembly into the PUC19 (Invitrogen) backbone. sgRNA sequence AGCCACGTATGACGTTATCA was cloned into pX330-U6-Chimeric_BB-CBh-hSpCas9. Cells were nucleofected with pX330 vector (RRID:Addgene_42230) containing the sgRNA and Cas9 and the repair template using the Nucleofector 4D (Lonza) with Kit SF and program DN-100. GFP^+ cells were sorted using a FACS AriaIII (BD Biosciences). In a second sorting round, GFP^+ cells were single cell sorted and tested for proper integration. Clone 1A5 was transduced with lenti pCW-Cas9 (RRID:Addgene_50661), puromycin selected (1 µg ml^{-1})

and subsequently single cell sorted based on GFP positivity and tested for inducible Cas9 expression. Clone 3E7 was used for the screen, which we called MUTZ3-EVI1-GFP.

Patient material

Samples of the selected patients presenting with AML were collected from the Erasmus MC Hematology Department biobank (Rotterdam, the Netherlands). The karyotype of AML patients used in this study was as follows; AML-1: 45,XX,inv(3)(q2?1q26),-7, AML-2: 45,XY,inv(3)(q22q26),-7 and AML-3: 45,XX,t(3;3)(q21;q26),-7. Leukemic blast cells were purified from bone marrow or blood by standard diagnostic procedures. All patients provided written informed consent in accordance with the Declaration of Helsinki. The Medical Ethical Committee of the Erasmus MC has approved usage of the patient rest material for this study.

Western Blotting

Cells were lysed in lysis buffer (20 mM Tris-HCL, 138 mM NaCl, 10 mM EDTA, 50 mM NaF, 1% Triton, 10% glycerol, 2 mM NA-vanadate) containing Complete protease inhibitors (CPI, Roche #4693159001). Protein levels were detected using antibodies against EVI1 (Cell Signaling, Cat# 2265, RRID:AB_561424), MYB (Millipore Cat# 05-175, RRID:AB_2148022), FLAG (Sigma-Aldrich Cat# F3165, RRID:AB_259529), B-Actin (Sigma-Aldrich Cat# A5441, RRID:AB_476744), GAPDH (Santa Cruz Biotechnology Cat# sc-25778, RRID:AB_10167668), CAS9 (Biolegend Cat# 844301, RRID:AB_2565570) or GATA2 (kind gift of E.H. Bresnick, Department of Cell and Regenerative Biology, Madison, WI). Proteins were visualized using the Odyssey infrared imaging system (Li-Cor).

Flow cytometric analysis

Cell Sorting was performed using the FACS Aria flow cytometer (BD Biosciences, RRID:SCR_013311) into a 96-well plate format or into batch culture. Flow Cytometric analysis on MUTZ3 cells was done with GFP/RFP or antibody stainings for CD34-PE-CY7 (BD Biosciences Cat# 348811, RRID:AB_2868855) and CD15-APC (Sony, #2215035) or CD15-BV510 (BioLegend Cat# 323028, RRID:AB_2563400). Intracellular stainings with EVI1 (Cell Signaling Technology Cat# 2256, RRID:AB_561017) or Rabbit (DA1E) mAb IgG XP® Isotype Control (Cell Signaling Technology Cat# 3900, RRID:AB_1550038) were performed using Foxp3/Transcription Factor Staining Buffer Set (00-5523-00, eBioscience). Cells were measured on a BD Canto or BD LSR II flow cytometer (BD Biosciences), and data was analyzed using FlowJo software (FlowJo, RRID:SCR_008520).

DNA pulldown

Nuclear lysates for pulldown experiments were prepared as described [45]. Oligo nucleotides for affinity purification were ordered as custom-synthesized oligos from Integrated DNA Technologies (IDT) (see Table S4). DNA pulldown was performed as described by Karemaker and Vermeulen with minor changes. Essentially, per DNA pulldown, 500 pmole of annealed oligos were diluted to 600 µL in DNA binding buffer (DBB: 1 M NaCl, 10 mM Tris pH 8.0, 1 mM EDTA, 0.05% NP40) and incubated with washed beads (10 uL Streptavidin Sepharose High performance bead slurry (GE Healthcare #17511301), washed once with PBS + 0.1% NP-40 and once with DBB) for 30 minutes at 4°C while rotating. After washing once with 1mL DBB and twice with 1 mL protein incubation buffer (PIB: 150 mM NaCl, 50 mM Tris pH 8.0, 0.25% NP40, 1 mM DTT with Complete protease inhibitors (CPI, Roche #4693159001)) the immobilized oligos on beads were combined with 500 µg nuclear extracts in a total volume of 600 µL PIB with 10 µg competitor DNA (5 µg poly-dIdC (Sigma #81349_500ug) and 5 µg poly-dAdt (Sigma #P0883_50UN)) and incubated for 90 minutes at 4°C while rotating. Beads were washed three times with 1mL PIB and twice with 1 mL PBS. To elute proteins from the oligo probes, beads were resuspended in 20 uL 1x western blot protein sample buffer and incubated at 95°C for 15 minutes while shaking. The beads were spun down and the eluate was loaded on a protein gel. A 40 µg nuclear extract sample was prepared directly from the nuclear lysate as input sample for western blot.

Peptide treatment of cells

MUTZ3, primary AMLs or $CD34^+$ cells were cultured in medium as described above, plus MYBMIM or control peptide TG3 at indicated concentrations. For measuring viability of MUTZ3 or primary AMLs, cells were seeded in an opaque colored 96-well plate at 15.000 cells/well in a total volume of 100 µl medium containing MYBMIM or control peptide TG3 at indicated concentrations (20 µM MYBMIM or control peptide TG3 for primary AMLs). Cell viability was assessed 72 hours after treatment using CellTiter-Glo cell viability assay according to manufacturer's protocol (Promega). Luminescence was measured on the Victor X3 plate reader (Perkin Elmer). Rescue experiments in MUTZ3 cells were performed by retroviral overexpression of murine pMY-FLAG-Evi1-IRES-GFP or an empty vector (EV) control. The pMY vectors were kind gifts of T. Sato [26] in which FLAG was inserted 5' of *Evi1*. *Evi1* or EV overexpressing cells were cultured in the presence of 20 µM MYBMIM for 48 hours. For colony cultures following peptide treatment, 2000 MUTZ3 cells or 500 $CD34^+$ cells were plated in MethoCult (StemCell technologies) with 100U/ml penicillin/streptomycin. For protein lysates and ChIP experiments cells were cultured containing 20 µM MYBMIM or control peptide TG3 and harvested after 48 hours of peptide treatment.

Fluorescence in situ hybridization (FISH)

FISH was performed and reported according to standard protocols based on the International System of Human Cytogenetics Nomenclature (2016) [46]. *MECOM* FISH was performed according to the manufacturer's protocol using the *MECOM* t(3;3); inv(3)(3q26) triple-color probe (Cytocell, LPH-036).

Genome editing

The sgRNAs (Table S1) were either cloned into pLentiV2_U6-IT-mPgk-iRFP720 (J.Z.) using BsmBI restriction sites, px330 using BbsI or were *in vitro* transcribed using the T7 promoter. Lentiviruses were prepared by transfecting 293T cells with lentiviral packaging constructs pSPAX2/pMdelta2.G and sgRNA cloned into pLentiV2_U6-IT-mPgk-iRFP720. Transfections were performed using Fugene 6 (Promega) according to manufacturer's protocol. For *in vitro* transcribed sgRNAs oligo's containing the T7 promoter, target sequence and the Tail annealing sequence were annealed, filled in and transcribed using the Hi-scribe T7 kit (NEB). Turbo DNAse (Invitrogen) was added and sgRNAs were cleaned up using RNA clean&concentrator kit (Zymo). Concentration of sgRNAs was estimated using Qubit (Invitrogen). RNP complexes were formed incubating sgRNA and Cas9 (IDT) for 20-30 at RT before nucleofection using the Neon (Thermofischer) with buffer R with settings 1500V, 20ms, 1 pulse for MUZT3 or 1350V, 10ms, 4 pulses for K562. Genomic DNA was extracted at indicated timepoint after transfection using Quick Extract buffer (Epicenter) PureLink Genomic DNA Mini Kit (Invitrogen) and checked for targeting by PCR using Q5 polymerase (NEB) or amplicon-sequencing.

Pooled sgRNA Enhancer scanning

To design a high-resolution sgRNA library for the enhancer scan, we considered all possible sgRNA target sites containing a canonical Cas9 PAM site (NGG) on both strands of the minimal 18 kb translocated region. sgRNAs containing a G in positions 1-3 of the 20nt target site were trimmed at this position to favor 20-, 19- or 18-mers (in this order of priority) containing a natural G at the 5'end as previously described [47]. For all other sgRNAs, a G was added to the 5'end (resulting in a 21-mer). Subsequently, all sgRNAs showing (1) a high number of target sites in the human genome (>5 with no mismatch, or >20 with 1 mismatch), (2) a BsmBI site (interfering with cloning), or (3) a polyA signal (interfering with packaging) were filtered out. In addition, we added a number of negative controls (82 sgRNAs targeting the *AAVS1* region) as well as positive controls (5 sgRNAs targeting *EVI1* as well as 313 sgRNAs covering 5 kb of the breakpoint in MUTZ3 cells). The final library of 3239 sgRNAs (Table S2) was synthesized with overhangs for PCR amplification and cloning as one oligo pool (Twist Bioscience) and cloned into the lentiviral vector sgETN (J.Z.) as previously described [47]. The pool of 3239 sgETN-sgRNAs was transduced in triplicate into MUTZ3-EVI1-GFP. For each replicate, a total of 120 million cells were infected with 3-4% transduction efficiency to ensure that each

sgRNA is represented predominantly as a single lentiviral integration in >1000 cells. After neomycin drug selection (1 mg ml^{-1}) for 7 days, T0 samples were obtained (5 million cells per replicate), and cells were subsequently cultured in the presence of 1 µg ml^{-1} doxycycline (Dox). Culture medium was exchanged every 2 days. After 5 days (T5) and 7 days (T7), about 1 million sgRNA-expressing (GFP^{low}) cells were sorted for each replicate using a FACS AriaII (BD Biosciences). Genomic DNA from T0, T5 and T7 samples was isolated by two rounds of phenol extraction using PhaseLock tubes (5PRIME), followed by isopropanol precipitation. Deep-sequencing libraries were generated by PCR amplification of sgRNA guide strands using primers that tag the product with standard Illumina adapters and a 4 bp sample barcode in a 2 step-PCR protocol. For each sorted sample, all DNA was used as template in multiple parallel 50-µl PCR reactions, each containing 250-500 ng template, 1x AmpliTaq Gold buffer, 0.2 mM of each dNTP, 2 mM $MgCl_2$, 0.3 µM of each primer and 1U AmpliTaq Gold (Invitrogen), which were run using the following cycling parameters: 95 °C for 10 min; 28 cycles of 95 °C for 30 s, 52 °C for 45 s and 72 °C for 30 s; 72 °C for 7 min. PCR products (367 bp) were combined for each sample and Ampure purified. For the T0 samples and a DNA-pool sample the amount of input DNA necessary to get a 1000x coverage was used as input in the PCRs. For the second PCR 10ng of input was used per PCR using the following cycling parameters: 95 °C for 10 min; 8 cycles of 95 °C for 30 s, 57 °C for 45 s and 72 °C for 30 s; 72 °C for 7 min. PCR products (448 bp) were combined for each sample and Ampure-purified. Libraries were sequenced equimolarly on an Illumina HiSeq 2500 (Illumina) by the Next Generation Sequencing Facility at Vienna BioCenter Core Facilities (VBCF), member of the Vienna BioCenter (VBC), Austria. Multiple experiments (different time points and sorted fractions) were sequenced simultaneously, each identified by a unique barcode. Sequencing data were processed by converting unaligned BAM files into FASTA using bam2fastx. Experiment-specific barcodes (positions 7-10) were extracted together with the sgRNA sequence (positions 31-) into a new FASTA file, which was subsequently reverse-complemented with seqtk seq. Next, the barcodes were used to demultiplex the FASTA file into experiment-specific files with ngs-tools split-by-barcode, using parameters –s 4 –d 1, i.e. barcode size 4 and maximum 1 mismatch. For each of these files, we counted the number of identical sgRNA sequences with fastx_collapser and we assigned them to their known identifiers. These counts were employed for downstream data analysis. To provide a sufficient baseline for detecting sgRNA enrichment in experimental samples, we aimed to acquire >1000 reads per sgRNA in the sequenced sgRNA pool to compensate for variation in sgRNA representation inherent in the pooled plasmid preparation or introduced by PCR biases. Reads were normalized to the total number of library-specific reads per lane for each condition. To ensure a proper sgRNA representation in the initial plasmid pool, we used a cutoff of more than 10% average reads/sgRNA sequenced in the Plasmid-Pool (resulting in passing of 3050 out of 3239 sgRNAs). Enrichment analyses were performed using MAGeCK [48].

ChIP sequencing

$H_3K_{27}Ac$ and p300 ChIP-seq data from the inv(3) cell line MOLM1 as well as p300 ChIP-seq data from MUTZ3 were previously generated by our group and are available at ArrayExpress E-MTAB-2224 [12]. $H_3K_{27}Ac$ (Abcam Cat# ab4729, RRID:AB_2118291) ChIPs were performed according to the standard ChIP protocol from Upstate. ChIP with antibodies direct against MYB (Millipore Cat# 05-175, RRID:AB_2148022) or p300 (Diagenode, #C15200211) were performed by first crosslinking for 45 minutes with DSG before formaldehyde crosslinking. ChIP samples were processed according to the Illumina TruSeq ChIP Sample Preparation Protocol (Illumina) or Diagenode Library V3 preparation protocol (Diagenode) and either sequenced single-end (1x 50 bp) on the HiSeq 2500 platform (Illumina) or paired-end (2x100 bp) on the Novaseq 6000 platform (Illumina). Briefly, reads were aligned to the human reference genome build hg19 with bowtie [49] for single-end runs and bowtie2 [50] for paired-end runs, and bigwig files were generated for visualization with bedtools genomecov [51] and UCSC bedGraphToBigWig [52]. Peaks were determined using the MACS2 program with default parameters [53]. The tracks were normalised per million reads (RPM) and visualized as genome browser profiles using the Fluff package [54].

Cut&Run

$H_3K_{27}Ac$ (Abcam Cat# ab4729, RRID:AB_2118291) Cut&Run libraries for the MUTZ3 bulk and sorted fragments were generated with an input of 200.000 cells. The protocol described by the Henikoff group was used to generate these tracks [55], using a 0.04% Digitonin buffer and with the addition of cOmplete, EDTA-free Protease Inhibitor Cocktail (Roche) and 1M Sodiumbutyrate (Sigma Aldrich) to all the buffers. Isolation was done according to the standard Phenol Chloroform protocol. Cut&Run samples were processed according to the protocol described by the Fazzio group [56] and sequenced paired-end (2x100 bp) on the Novaseq 6000 platform (Illumina). Reads were aligned similarly to ChIP-seq.

ATAC sequencing

Open chromatin regions were mapped by the ATAC-seq method as described [57] with a modification in the lysis buffer (0.30 M sucrose, 10 mM Tris pH 7.5, 60 mM KCl, 15 mM NaCl, 5 mM MgCl2, 0.1 mM EGTA, 0.1% NP40, 0.15 mM Spermine, 0.5 mM Spermidine, 2 mM 6AA) to reduce mitochondrial DNA contamination. ATAC-seq samples were sequenced paired-end (2x 50 bp) on the HiSeq 2500 platform (Illumina) and aligned against the human genome (hg19) with bowtie2, allowing for a maximum 2000 bp insert size. Mitochondrial reads and fragments with mapping quality below 10 were removed.

RNA sequencing

RNA was isolated either using Trizol or the Qiagen Allprep DNA/RNA kit and protocol (Qiagen, #80204). cDNA synthesis was done using the SuperScript II Reverse Transcriptase

kit (Invitrogen). Quantitative real-time PCR was performed by using primers (Table S4) as described previously [15] on the 7500 Fast Real-time PCR System (Applied Biosystems). For RNA sequencing, sample libraries were prepped using 500 ng of input RNA according to the KAPA RNA HyperPrep Kit with RiboErase (HMR) (Roche) using Unique Dual Index adapters (Integrated DNA Technologies, Inc.). Amplified sample libraries were paired-end sequenced (2x100 bp) on the Novaseq 6000 platform (Illumina) and aligned against the human genome (hg19) using STAR version 2.5.4b. Salmon [58] was used to quantify expression of individual transcripts, which were subsequently aggregated to estimate gene-level abundances with tximport [59]. Human gene annotation derived from RefSeq [60] was downloaded from UCSC [61] (RefGene) as a GTF file. Transcript-level abundances were normalized to transcripts per million (TPM) for visualization.

Amplicon sequencing

For amplicon sequencing we used a PCR-based NGS library preparation method in combination with the TruSeq Custom Amplicon index kit (Illumina). The first PCR for target selection (Table S4) was performed using Q5 polymerase (NEB), the second nested PCR, to add the index-adapters, with KAPA HiFi HotStart Ready mix (KapaBiosystems). Libraries were sequenced paired-end (2x 250 bp) on the MiSeq platform (Illumina). Reads were trimmed with trimgalore (Trim Galore, RRID:SCR_011847) to remove low-quality bases and adapters, and subsequently aligned to the human reference genome build hg19 with BBMap (BBmap, RRID:SCR_016965) allowing for 1000 bp indels. Mutations introduced by genome editing were analysed and visualised using CRISPResso2 [62]. Mutated sequences consisting of up to 5% of sequenced reads were next analysed for differential binding with CIS- BP [63].

AUTHOR CONTRIBUTIONS

L.S. and R.D. conceived and designed the experiments. J.Z. co-designed and supervised the enhancer screen and provided critical resources. L.S., S.O., A.E., M.F., M.H., A.A.V., D.P., S.v.H., C.E., T.G. and E.B. performed experiments. R.M.L. was responsible for all bioinformatic data processing and performed the majority of the bioinformatic analyses, R.H. and L.S. performed certain bioinformatic analyses. F.G.K., D.R.M., E.H.B. and A.K. provided important resources for the study. L.S., R.D., S.O. and R.M.L wrote the manuscript with input from all authors.

My contributions to this work were: processing and analysis of all high throughput sequencing data (RNA-seq, ATAC-seq, ChIP-seq, Cut&Run); analysis of the CRISPR enhancer scan; data management and upload; interpretation of the results and writing of the manuscript.

ACKNOWLEDGMENTS

We thank our colleague Michael Vermeulen and the Bioptics Facility at IMP for flow cytometric sorting. We are thankful to Tobias Neumann for help with the design of the enhancer scanning strategy as well as to the Zuber group at the IMP for their help with the enhancer scanning CRISPR/Cas9 experiments. We thank the Vermeulen group at the Radboud Institute for Molecular Life Sciences for assistance in the DNA-pulldown experiments. Furthermore, we acknowledge Berna Beverloo and the department of Clinical Genetics for the FISH analysis, and colleagues from the bone marrow transplantation group and the molecular diagnostics laboratory of the Department of Hematology for storage of samples and molecular analysis of the leukemia cells. This work was funded by a fellowship from the Daniel den Hoed, Erasmus MC Foundation (L.S.), the Koningin Wilhelmina Fonds grant from the Dutch Cancer Society (R.D., R.M., S.O., and T.G.), the National Institutes of Health grant R01 DK68634 (E.H.B), Carbone Cancer Center P30 CA014520 (E.H.B), the National Institutes of Health T32 HL07899 (D.R.M.) and FWF-SFB grant F4710 of the Austrian Science Fund (J.Z). Research at the IMP is generously supported by Boehringer Ingelheim and the Austrian Research Promotion Agency (Headquarter grant FFG-852936).

CONFLICT OF INTEREST DISCLOSURE

A patent application related to MYBMIM has been submitted by A.K. to the U.S. Patent and Trademark Office entitled ‘‘Agents and methods for treating CREB binding protein-dependent cancers’’ (application PCT/US2017/059579). A.K. received personal fees from Novartis and from Rgenta during the conduct of the study.

REFERENCES

1 Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* **153**, 17-37, doi:10.1016/j.cell.2013.03.002 (2013).

2 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).

3 Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195, doi:10.1126/science.1222794 (2012).

4 Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat Rev Genet* **17**, 93-108, doi:10.1038/nrg.2015.17 (2016).

5 Zhu, H. *et al.* Candidate Cancer Driver Mutations in Distal Regulatory Elements and Long-Range Chromatin Interaction Networks. *Mol Cell* **77**, 1307-1321 e1310, doi:10.1016/j.molcel.2019.12.027 (2020).

6 Rahman, S. & Mansour, M. R. The role of noncoding mutations in blood cancers. *Dis Model Mech* **12**, doi:10.1242/dmm.041988 (2019).

7 Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat Genet* **46**, 1258-1263, doi:10.1038/ng.3141 (2014).

8 Bresnick, E. H. & Johnson, K. D. Blood disease-causing and -suppressing transcriptional enhancers: general principles and GATA2 mechanisms. *Blood Adv* **3**, 2045-2056, doi:10.1182/bloodadvances.2019000378 (2019).

9 Mansour, M. R. *et al.* Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373-1377, doi:10.1126/science.1259037 (2014).

10 Herranz, D. *et al.* A NOTCH1-driven MYC enhancer promotes T cell development, transformation and acute lymphoblastic leukemia. *Nat Med* **20**, 1130-1137, doi:10.1038/nm.3665 (2014).

11 Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454-1458, doi:10.1126/science.aad9024 (2016).

12 Gröschel, S. *et al.* A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* **157**, 369-381, doi:S0092-8674(14)00218-9 [pii] 10.1016/j.cell.2014.02.019 (2014).

13 Northcott, P. A. *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428-434, doi:10.1038/nature13379 (2014).

14 Affer, M. *et al.* Promiscuous MYC locus rearrangements hijack enhancers but mostly super-enhancers to dysregulate MYC expression in multiple myeloma. *Leukemia* **28**, 1725-1735, doi:10.1038/leu.2014.70 (2014).

15 Ottema, S. *et al.* Atypical 3q26/MECOM rearrangements genocopy inv(3)/t(3;3) in acute myeloid leukemia. *Blood* **136**, 224-234, doi:10.1182/blood.2019003701 (2020).

16 Barjesteh van Waalwijk van Doorn-Khosrovani, S. *et al.* High EVI1 expression predicts poor survival in acute myeloid leukemia: a study of 319 de novo AML patients. *Blood* **101**, 837-845, doi:10.1182/blood-2002-05-1459 (2003).

17 Lugthart, S. *et al.* Clinical, Molecular, and Prognostic Significance of WHO Type inv(3)(q21q26.2)/t(3;3)(q21;q26.2) and Various Other 3q Abnormalities in Acute Myeloid Leukemia. *Journal of Clinical Oncology* **28**, 3890-3898, doi:10.1200/jco.2010.29.2771 (2010).

18 Lugthart, S. *et al.* High EVI1 levels predict adverse outcome in acute myeloid leukemia: prevalence of EVI1 overexpression and chromosome 3q26 abnormalities underestimated. *Blood* **111**, 4329-4337, doi:10.1182/blood-2007-10-119230 (2008).

19 Morishita, K. *et al.* Activation of EVI1 gene expression in human acute myelogenous leukemias by translocations spanning 300-400 kilobases on chromosome band 3q26. *Proceedings of the National Academy of Sciences* **89**, 3937-3941, doi:10.1073/pnas.89.9.3937 (1992).

20 Yamazaki, H. *et al.* A remote GATA2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating EVI1 expression. *Cancer Cell* **25**, 415-427, doi:S1535-6108(14)00076-2 [pii] 10.1016/j.ccr.2014.02.008 (2014).

21 Grass, J. A. *et al.* Distinct functions of dispersed GATA factor complexes at an endogenous gene locus. *Mol Cell Biol* **26**, 7056-7067, doi:10.1128/MCB.01033-06 (2006).

22 Johnson, K. D. *et al.* Cis-regulatory mechanisms governing stem and progenitor cell transitions. *Sci Adv* **1**, e1500503, doi:10.1126/sciadv.1500503 (2015).

23 Wilson, N. K. *et al.* Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* **7**, 532-544, doi:10.1016/j.stem.2010.07.016 (2010).

24 Beck, D. *et al.* Genome-wide analysis of transcriptional regulators in human HSPCs reveals a densely interconnected network of coding and noncoding genes. *Blood* **122**, e12-22, doi:10.1182/blood-2013-03-490425 (2013).

25 Ramaswamy, K. *et al.* Peptidomimetic blockade of MYB in acute myeloid leukemia. *Nat Commun* **9**, 110, doi:10.1038/s41467-017-02618-6 (2018).

26 Yoshimi, A. *et al.* Evi1 represses PTEN expression and activates PI3K/AKT/mTOR via interactions with polycomb proteins. *Blood* **117**, 3617-3628, doi:10.1182/blood-2009-12-261602 (2011).

27 Loven, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320-334, doi:S0092-8674(13)00393-0 [pii] 10.1016/j.cell.2013.03.036 (2013).

28 Drier, Y. *et al.* An oncogenic MYB feedback loop drives alternate cell fates in adenoid cystic carcinoma. *Nat Genet* **48**, 265-272, doi:10.1038/ng.3502 (2016).

29 Sakamoto, H. *et al.* Proper levels of c-Myb are discretely defined at distinct steps of hematopoietic cell development. *Blood* **108**, 896-903, doi:10.1182/blood-2005-09-3846 (2006).

30 Beug, H., von Kirchbach, A., Doderlein, G., Conscience, J. F. & Graf, T. Chicken hematopoietic cells transformed by seven strains of defective avian leukemia viruses display three distinct phenotypes of differentiation. *Cell* **18**, 375-390, doi:10.1016/0092-8674(79)90057-6 (1979).

31 Weston, K. & Bishop, J. M. Transcriptional activation by the v-myb oncogene and its cellular progenitor, c-myb. *Cell* **58**, 85-93, doi:10.1016/0092-8674(89)90405-4 (1989).

32 Rahman, S. *et al.* Activation of the LMO2 oncogene through a somatically acquired neomorphic promoter in T-cell acute lymphoblastic leukemia. *Blood* **129**, 3221-3226, doi:10.1182/blood-2016-09-742148 (2017).

33 Nguyen, N. *et al.* Myb expression is critical for myeloid leukemia development induced by Setbp1 activation. *Oncotarget* **7**, 86300-86312, doi:10.18632/oncotarget.13383 (2016).

34 Ramsay, R. G. & Gonda, T. J. MYB function in normal and cancer cells. *Nat Rev Cancer* **8**, 523-534, doi:10.1038/nrc2439 (2008).

35 Zuber, J. *et al.* An integrated approach to dissecting oncogene addiction implicates a Myb-coordinated self-renewal program as essential for leukemia maintenance. *Genes Dev* **25**, 1628-1640, doi:10.1101/gad.17269211 (2011).

36 Kasper, L. H. *et al.* A transcription-factor-binding surface of coactivator p300 is required for haematopoiesis. *Nature* **419**, 738-743, doi:10.1038/nature01062 (2002).

37 Sandberg, M. L. *et al.* c-Myb and p300 regulate hematopoietic stem cell proliferation and differentiation. *Dev Cell* **8**, 153-166, doi:10.1016/j.devcel.2004.12.015 (2005).

38 Best, J. L. *et al.* Identification of small-molecule antagonists that inhibit an activator: coactivator interaction. *Proc Natl Acad Sci U S A* **101**, 17622-17627, doi:10.1073/pnas.0406374101 (2004).

39 Uttarkar, S. *et al.* Naphthol AS-E Phosphate Inhibits the Activity of the Transcription Factor Myb by Blocking the Interaction with the KIX Domain of the Coactivator p300. *Mol Cancer Ther* **14**, 1276-1285, doi:10.1158/1535-7163.MCT-14-0662 (2015).

40 Uttarkar, S. *et al.* Small-Molecule Disruption of the Myb/p300 Cooperation Targets Acute Myeloid Leukemia Cells. *Mol Cancer Ther* **15**, 2905-2915, doi:10.1158/1535-7163.MCT-16-0185 (2016).

41 Walf-Vorderwulbecke, V. *et al.* Targeting acute myeloid leukemia by drug-induced c-MYB degradation. *Leukemia* **32**, 882-889, doi:10.1038/leu.2017.317 (2018).

42 Takao, S. *et al.* Convergent organization of aberrant MYB complex controls oncogenic gene expression in acute myeloid leukemia. *Elife* **10**, doi:10.7554/eLife.65905 (2021).

43 Roe, J. S., Mercan, F., Rivera, K., Pappin, D. J. & Vakoc, C. R. BET Bromodomain Inhibition Suppresses the Function of Hematopoietic Transcription Factors in Acute Myeloid Leukemia. *Mol Cell* **58**, 1028-1039, doi:10.1016/j.molcel.2015.04.011 (2015).

44 Diffner, E. *et al.* Activity of a heptad of transcription factors is associated with stem cell programs and clinical outcome in acute myeloid leukemia. *Blood* **121**, 2289-2300, doi:10.1182/blood-2012-07-446120 (2013).

45 Karemaker, I. D. & Vermeulen, M. ZBTB2 reads unmethylated CpG island promoters and regulates embryonic stem cell differentiation. *EMBO Rep* **19**, doi:10.15252/embr.201744993 (2018).

46 McGowan-Jordan, J., Simons, A. & Schmid, M. *ISCN : an international system for human cytogenomic nomenclature (2016)*. (Basel ; New York : Karger, 2016).

47 Michlits, G. *et al.* Multilayered VBC score predicts sgRNAs that efficiently generate loss-of-function alleles. *Nat Methods* **17**, 708-716, doi:10.1038/s41592-020-0850-8 (2020).

48 Li, W. *et al.* MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biology* **15**, 554, doi:10.1186/s13059-014-0554-4 (2014).

49 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25, doi:10.1186/gb-2009-10-3-r25 (2009).

50 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

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

52 Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204-2207, doi:10.1093/bioinformatics/btq351 (2010).

53 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).

54 Georgiou, G. & van Heeringen, S. J. fluff: exploratory analysis and visualization of high-throughput sequencing data. *PeerJ* **4**, e2209, doi:10.7717/peerj.2209 (2016).

55 Skene, P. J., Henikoff, J. G. & Henikoff, S. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nat Protoc* **13**, 1006-1019, doi:10.1038/nprot.2018.015 (2018).

56 Hainer, S. J., Boskovic, A., McCannell, K. N., Rando, O. J. & Fazzio, T. G. Profiling of Pluripotency Factors in Single Cells and Early Embryos. *Cell* **177**, 1319-1329 e1311, doi:10.1016/j.cell.2019.03.014 (2019).

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

58 Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419, doi:10.1038/nmeth.4197 (2017).

59 Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521, doi:10.12688/f1000research.7563.2 (2015).

60 O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745, doi:10.1093/nar/gkv1189 (2016).

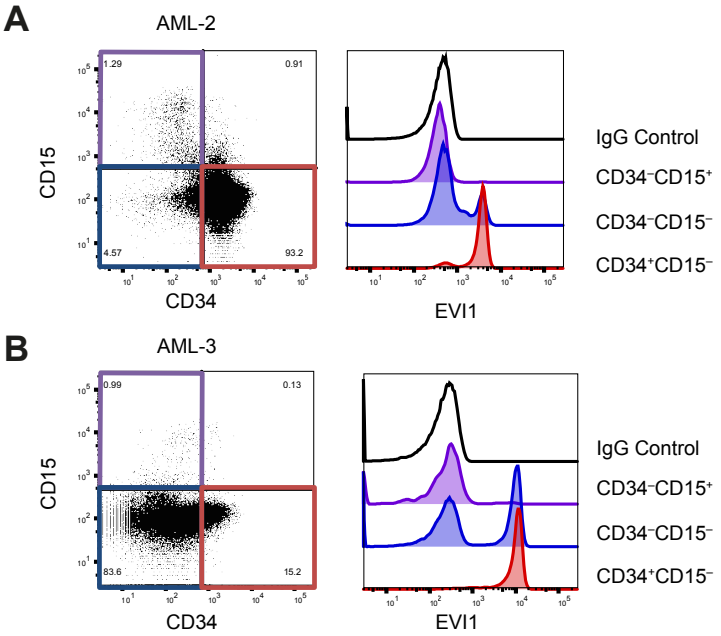
61 Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**, D493-496, doi:10.1093/nar/gkh103 (2004).

62 Clement, K. *et al.* CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat Biotechnol* **37**, 224-226, doi:10.1038/s41587-019-0032-3 (2019).

63 Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443, doi:10.1016/j.cell.2014.08.009 (2014).

SUPPLEMENTARY INFORMATION

SUPPLEMENTARY FIGURES



C

FISH for t(3;3); inv(3)(3q26)

	$CD34^+CD15^-$	$CD34^-CD15^-$	$CD34^-CD15^+$
Mutz3	100%	100%	100%
AML-3	98%	97%	82%

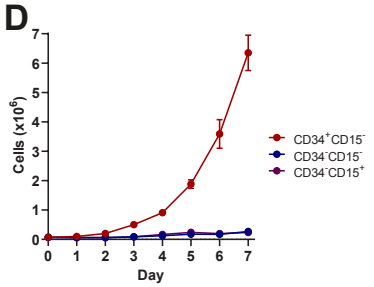


Figure S1. Expression of *EVI1* in inv(3)/t(3;3) AML is reversible. A. Flow cytometric analysis of CD34- and CD15-stained inv(3;3) primary AML cells (AML-2) (left) and intracellular EVI1 in the gated fractions (right). **B.** Flow cytometric analysis of CD34- and CD15-stained inv(3;3) primary AML cells (AML-3) (left) and intracellular EVI1 staining in the gated fractions (right). **C.** Percentage of cells which were found positive for *EVI1/3q26* rearrangements determined by three-colored FISH. **D.** Line graph showing numbers of MUTZ3 cells sorted into the indicated fractions and cultured for seven days. Error bars show standard deviation across three plates.

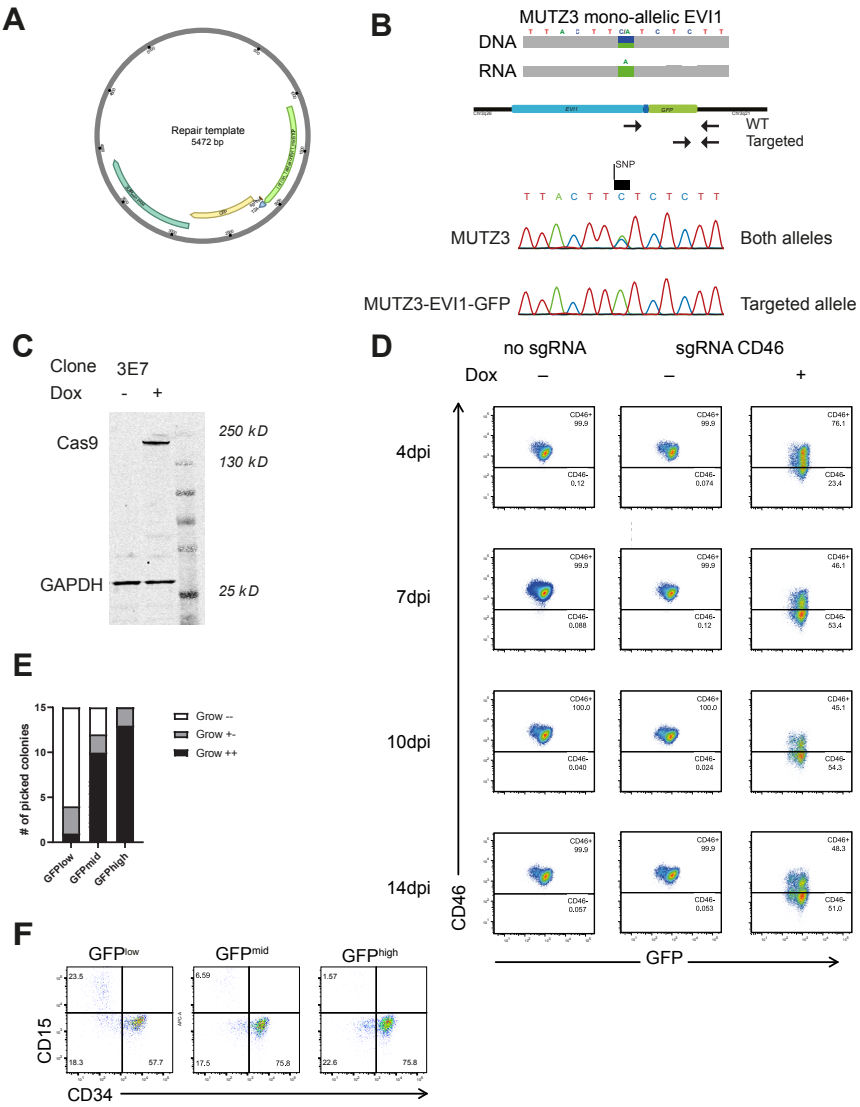


Figure S2. Generation of an EVI1-GFP inv(3) AML model. A. Repair template consisting of a PUC19 backbone with a homology arm of the intron and last exon of *EVI1* minus the STOP, a T2A site and GFP and the second homology arm. The PAM sequence of sgRNA was omitted. **B.** Mono-allelic expression of *EVI1* in MUTZ3 cells based on SNP differences (top). PCR strategy (middle) for Sanger sequencing of genomic DNA of MUTZ3 WT and MUTZ3-EVI1-GFP (bottom). **C.** Western blot showing Dox-inducible Cas9 protein expression in MUTZ3-EVI1-GFP. Cas9 was induced with 1ug/ml Dox for 48h. **D.** Testing tightness of the system by sgRNA-mediated knockout of the cell surface marker CD46. Cells were transduced and followed up by flow cytometric analysis for two weeks at indicated days post infection (dpi). Without Dox no knockout of CD46 is detected, whereas upon Dox exposure a strong effect on CD46 levels was observed. **E.** Growth in liquid cultures of fifteen colonies picked from methylcellulose (from experiment Figure 2H) from each sorted fraction. For each well, growth was defined as no growth (--), slow growth (+/-) or normal growth (++). **F.** Flow cytometric analysis of cells of sorted fractions 12 days after plating in methylcellulose stained with CD34 and CD15.

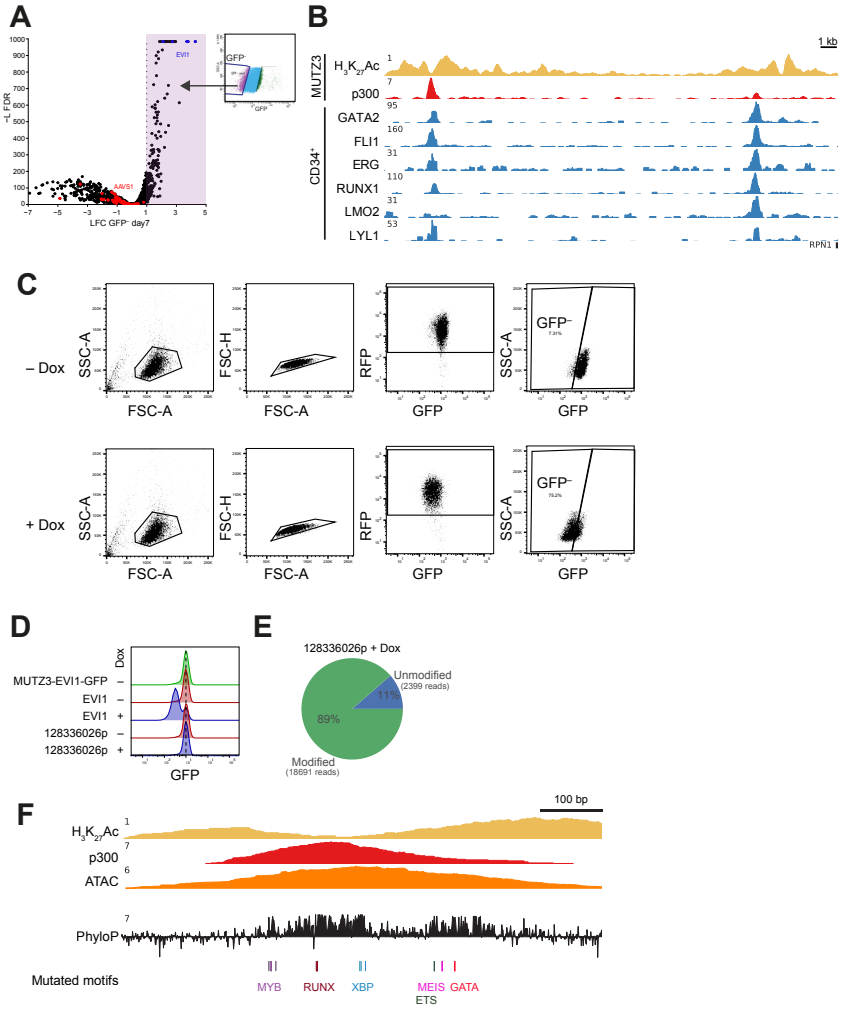


Figure S3. Unbiased CRISPR/Cas9 enhancer scan reveals one 1 kb region to be essential for *EVI1* activation. **A.** Volcano plot of enrichment of sgRNAs in sorted GFP^{low} fractions at day 7, based on three independent experiments. For every sgRNA detected in the GFP^{low} fractions, the log2fold change (LFC) compared to –Dox values was plotted as a dot. Five sgRNAs targeting *EVI1* were added to the sgRNA library as positive controls and are indicated in blue, sgRNAs targeting the AAVS1 region were added as negative controls indicated in red. The area highlighted in purple shows sgRNAs enriched at least two-fold (L2FC>1). **B.** Genome profile of the 18 kb translocated region with ChIP-seq showing binding of $H_3K_{27}Ac$, p300 and MYB in MUTZ3 cells and ChIP-seq of heptad transcription factors in CD34$^+$ cells (24). **C.** Flow cytometry gating strategy for validation experiments. A lentiviral vector containing sgRNA8 was transduced, Dox was added (+/- 4 days after transduction) and the cells were analyzed by flow cytometry at day 7 after Dox admission. **D.** Flowcytometric analysis of MUTZ3-EVI1-GFP cells upon sgRNA treatment. GFP signal shifts are shown upon transduction with lentivirus containing sgRNAs targeting *EVI1* or a sgRNA targeting a region (Chr3:128336026) which was not enriched in the enhancer scan. Cells were analyzed by flow cytometry 7 days after Dox induction of Cas9. **E.** Editing frequency of the sgRNA 128336026p. Modified reads exhibited variations with respect to the reference human sequence. **F.** Zoom-in of the -110 kb *GATA2* enhancer showing binding of $H_3K_{27}Ac$, p300 and open chromatin (ATAC), a track containing conservation scoring by phyloP (phylogenetic p-values) from the PHAST package (http://compgen.bscb.cornell.edu/phast/) and the mutations affecting motifs for known transcription factors in the GFP^{neg} fractions.

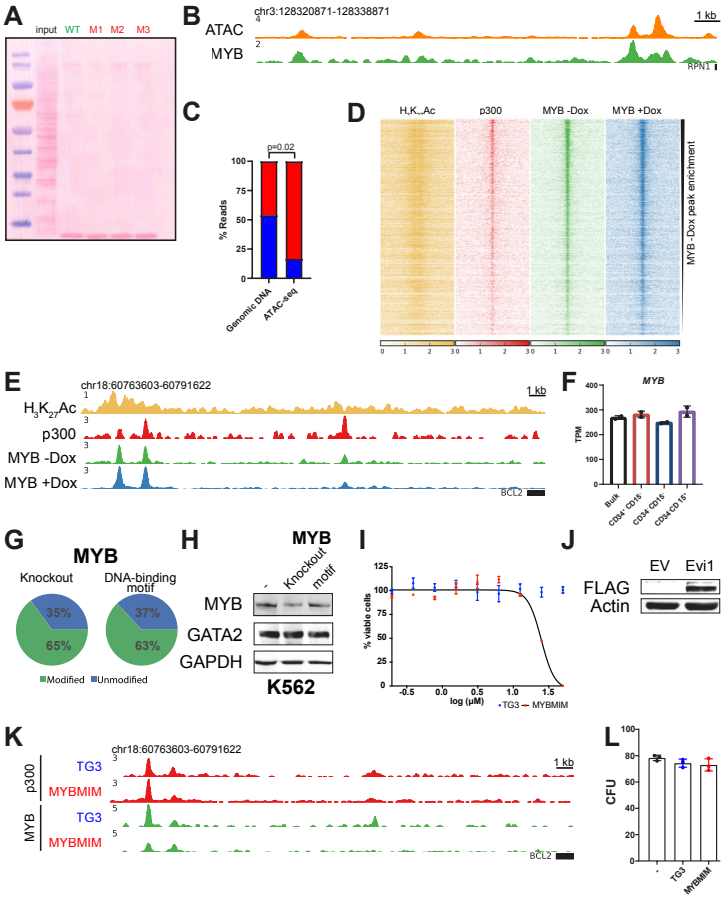


Figure S4. A MYB binding motif is essential for *EVI1* rather than for *GATA2* transcription. A. Ponceau red staining showing equal loading of protein of MUTZ3 protein lysates after DNA pulldown with immobilized WT, M1, M2 or M3 sequences. **B.** ATAC-seq and MYB ChIP-seq profiles of the *GATA2* enhancer region in CD34$^+$ Bone Marrow and CD34$^+$ Cord Blood cells, respectively. **C.** Bar plot showing allelic bias towards the translocated allele for ATAC-seq analysis based on a SNP (rs553101013). P-values were calculated using a $\chi 2$ test. **D.** Heatmap showing genome-wide MYB binding peaks as determined by the MACS2 peak calling algorithm, sorted for fold enrichment of MYB without Dox exposure. $H_3K_{27}Ac$ and p300 binding at those sites are shown as well. **E.** Pattern of $H_3K_{27}Ac$, p300 ChIP-seq in MUTZ3 cells at the BCL2 enhancer as well as MYB binding in MUTZ3-EVI1-GFP cells transduced with sgRNA8 with (+) or without (-) Dox induction of Cas9 for 7 days. **F.** Bar plot showing relative expression of *MYB* in Transcripts Per Million (TPM) in sorted fractions of MUTZ3 cells. Error bars represent standard deviation of two biological replicates. **G.** Editing frequency of K562 cells nucleofected with pX330-sgMYB.30 (Amplicon-seq_MYB.30) or pX330-sgRNA8 at day5 after nucleofection (Amplicon-seq_set1). Modified reads exhibited variations with respect to the reference human sequence. **H.** Western blot for MYB and GATA2 in K562 upon sgRNA-mediated MYB knockout (MYB.30) or MYB motif mutation (sgRNA8) at day 5 after nucleofection. GAPDH was used as loading control. **I.** MUTZ3 cells were treated with either TG3 (blue) or MYBMIM (red) at indicated concentrations and cell viability was determined by CellTiter-Glo three days after plating the cells in triplicate. **J.** Western blot using FLAG-specific antibody of MUTZ3 nuclear extracts with pMY-FLAG-Evi1-IRES-GFP (Evi1) or empty vector (EV). After viral transduction GFP expressing cells were sorted and Evi1 expression was confirmed by Western Blot. Actin was used as loading control. **K.** p300 and MYB ChIP-seq profiles of the BCL2 region in MUTZ3 cells treated with either 20 uM TG3 or MYBMIM for 48 h. **L.** Colony forming units (CFU) of CD34$^+$ Cord Blood cells. Cells were cultured without peptide or treated with 20uM TG3 or MYBMIM for two days, and subsequently plated in methylcellulose. Error bars show standard deviation across three plates.

CHAPTER 6

Allele-specific expression of GATA2 due to epigenetic dysregulation in CEBPA double mutant AML

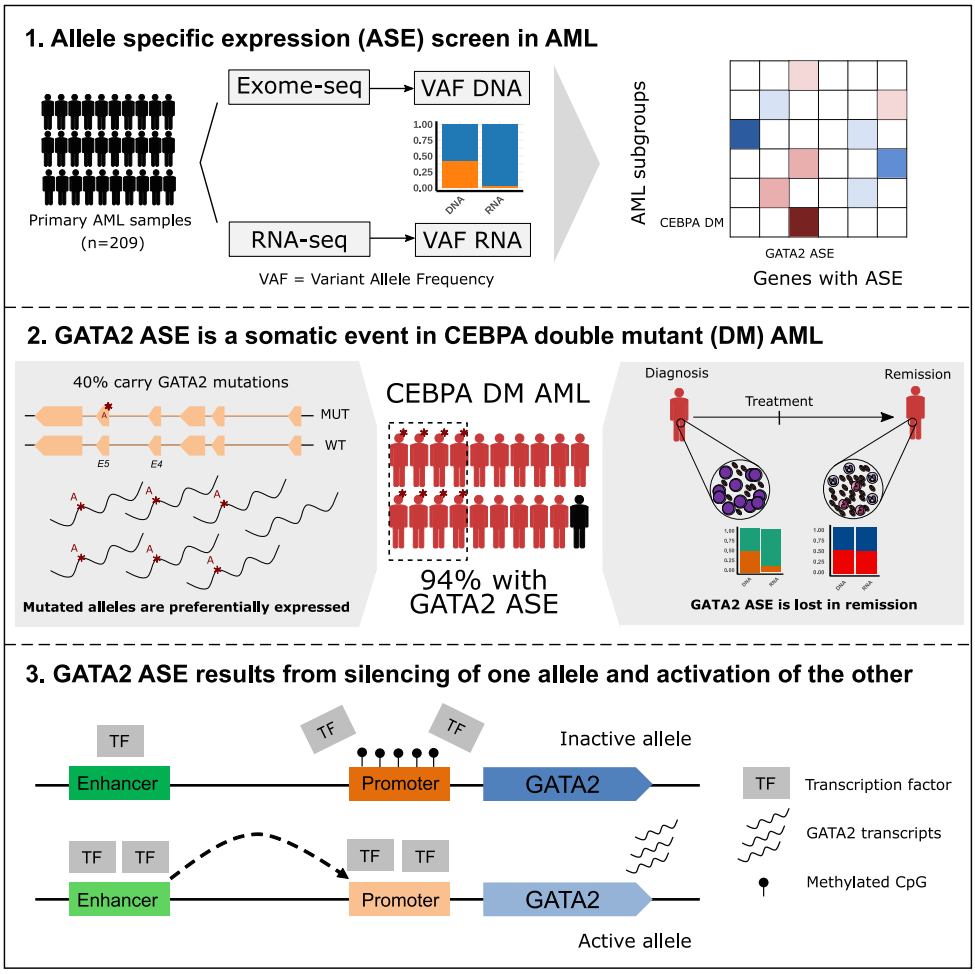
Roger Mulet-Lazaro[1,2], Stanley van Herk[1,2], Claudia Erpelinck[1,2], Eric Bindels[1], Mathijs A. Sanders[1], Carlo Vermeulen[2,4], Ivo Renkens[2,4], Peter Valk[1], Ari M. Melnick[3], Jeroen de Ridder[2,4], Michael Rehli[5,6], Claudia Gebhard[5,6], Ruud Delwel[1,2*], and Bas J. Wouters[1,2*]

[1]Department of Hematology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands
[2]Oncode Institute, Utrecht, The Netherlands
[3]Division of Hematology/Oncology, Department of Medicine, Weill Cornell Medical College, Cornell University, New York, United States
[4]Center for Molecular Medicine, University Medical Center, Utrecht University, Utrecht, the Netherlands
[5]Department of Internal Medicine III, University Hospital Regensburg, 93053, Regensburg, Germany
[6]Regensburg Center for Interventional Immunology (RCI), University Regensburg and University Medical Center Regensburg, 93053, Regensburg, Germany.

*These authors share senior authorship

Running title: GATA2 allele-specific expression in CEBPA DM AML

Blood. 2021 Jul 15;138(2):160-177. doi: 10.1182/blood.2020009244.
PMID: 33831168


1. Allele specific expression (ASE) screen in AML
Exome-seq
VAF DNA
RNA-seq
VAF RNA
Primary AML samples
(n=209)
VAF = Variant Allele Frequency
AML subgroups
CEBPA DM
GATA2 ASE
Genes with ASE
2. GATA2 ASE is a somatic event in CEBPA double mutant (DM) AML
40% carry GATA2 mutations
MUT
WT
E5
E4
Mutated alleles are preferentially expressed
CEBPA DM AML
94% with
GATA2 ASE
Diagnosis
Remission
Treatment
GATA2 ASE is lost in remission
3. GATA2 ASE results from silencing of one allele and activation of the other
TF
Enhancer
Promoter
GATA2
Inactive allele
Active allele
Transcription factor
GATA2 transcripts
Methylated CpG

ABSTRACT

Transcriptional deregulation is a central event in the development of acute myeloid leukemia (AML). To identify potential disturbances in gene regulation, we conducted an unbiased screen of allele-specific expression (ASE) in 209 AML cases. The gene encoding GATA binding protein 2 (*GATA2*) displayed ASE more often than any other myeloid or cancer-related gene. *GATA2* ASE was strongly associated with *CEBPA* double mutations (*CEBPA* DM), with 95% of cases presenting *GATA2* ASE. In *CEBPA* DM AML with *GATA2* mutations, the mutated allele was preferentially expressed. We found that *GATA2* ASE is a somatic event lost in complete remission, supporting the notion that it plays a role in *CEBPA* DM AML. Acquisition of *GATA2* ASE involved silencing of one allele via promoter methylation and concurrent overactivation of the other allele, thereby preserving expression levels. Notably, promoter methylation was also lost in remission together with *GATA2* ASE. In summary, we propose that *GATA2* ASE is acquired by epigenetic mechanisms and is a prerequisite for the development of AML with *CEBPA* DM. This finding constitutes a novel example of an epigenetic hit cooperating with a genetic hit in the pathogenesis of AML.

KEY POINTS

- *GATA2* ASE is a somatic event strongly associated with CEBPA double mutations in AML
- *GATA2* ASE results from silencing of one allele by promoter methylation and overactivation of a super-enhancer in the other allele

INTRODUCTION

Transcriptional deregulation is a central event in cancer development[1]. In acute myeloid leukemia (AML), most driver mutations occur in genes related to transcription, RNA splicing, chromatin regulation and/or DNA methylation[2]. In addition to mutations in protein-coding genes, alterations involving *cis*-regulatory elements play a critical role in aberrant gene expression in AML[3]. Examples include aberrant expression of *EVI1* thought translocation of the distal *GATA2* super-enhancer in AML with 3q26 aberrations[4] or focal amplification of distal *MYC* enhancers in AML with copy number changes in 8q24[5]. Other mechanisms identified in other malignancies include DNA alterations in *cis*-regulatory regions[6] and changes in binding sites for CTCF and cohesin[7]. Finally, in the absence of sequence variation, DNA methylation can modify gene expression, either directly by inducing promoter silencing[8] or by preventing CTCF binding[9].

Alterations in *cis*-regulatory regions usually affect a single DNA copy, leading to unbalanced expression of each allele controlled by these regulatory regions. For example, the gain of a super-enhancer selectively increases gene expression only in the allele where the new super-enhancer is created[10]. This phenomenon, termed allele-specific expression (ASE), can therefore serve as a tell-tale marker for *cis*-regulatory variation[11]. Besides acting as a surrogate marker, ASE can directly play a pathogenic role, e.g. by haploinsufficiency or preferential expression of a mutated protein[12]. In addition, ASE of specific genes may be associated with increased risk of cancer development[13] or progression[14], as shown for colon cancer[15], breast cancer and ovarian cancer[16] .

Extensive data focusing on the occurrence and relevance of ASE in AML are lacking. Here, we carried out a systematic study of genes with aberrant ASE in AML to uncover aberrantly expressed genes caused by abnormalities in *cis*-regulatory elements. To this end, we generated whole exome sequencing (WES) and RNA-seq data in a large representative cohort of AML patients and identified genes that recurrently exhibit ASE. Among those, *GATA2* stood out prominently and exhibited a strong association with *CEBPA* double mutant. A multi-omics analysis of the *GATA2* regulatory region showed that ASE is a result of concomitant promoter methylation on one allele and compensatory enhancer activation on the other allele.

METHODS

Allele-specific expression

To discriminate expression from different alleles, whole exome sequencing (WES) and RNA-seq data were integrated using an in-house python script. First, single nucleotide variants (SNVs) were detected on the WES data and, secondly, allele-specific read counts at every SNV were computed in both WES and RNA-seq data. SNVs with fewer than 9 WES reads or 5 RNA-seq reads were excluded. Information was aggregated over all the SNVs in a gene and ASE was determined with False Discovery Rate (FDR) < 0.05 in a $\chi 2$ test and RNA variant allele frequency < 0.35 (Figure 1). VAF below 0.1 was defined as the threshold for monoallelic expression. After the initial exploratory screen, a targeted, manually curated analysis was conducted on *GATA2* to identify cases missed by the automated pipeline -- ASE was defined only by RNA minor allele frequency < 0.35 for SNVs with more than 20 reads.

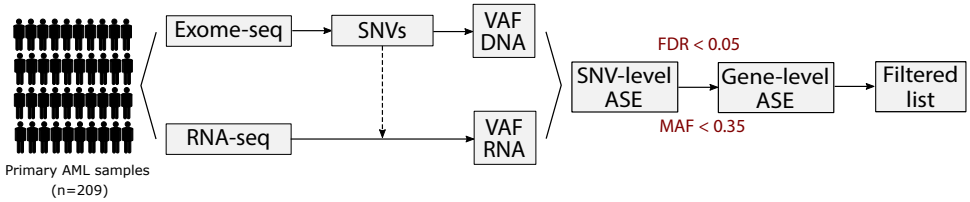


Figure 1. Schematic representation of the automated pipeline for allele specific expression detection. Raw reads were aligned by *STAR* (RNA-seq) or *bwa* (Exome-seq). Single nucleotide variants (SNVs) were called with an ensemble of programs and annotated based on function, population frequency and NGS statistics. This allowed the subsequent filtering of variants that were both real and informative. For every SNV, the variant allele frequency (VAF) at the DNA and RNA level was computed, and SNV information was aggregated at gene level. Finally, allele specific expression (ASE) was determined based on frequency of the minor allele (MAF) < 0.35 and false discovery rate (FDR) < 0.05 in a chi-square test.

Statistical association between mutations and genes with ASE

We calculated the statistical association between every possible pair of mutated genes and genes with ASE based on the co-occurrence of these two events in the patient cohort, using a Fisher's exact test. For descriptive statistics and hypotheses tests involving clinical variables, the R package *Atable*[17] was used with customized settings and functions.

Methylation analyses

Methylation analysis of the *GATA2* locus were conducted using ERRBS data previously published by our group[18] and bisulfite amplicon sequencing. Raw aligned reads and methylated base calls for CpGs were imported, filtered and normalized with the package *methylKit*[19] (v1.13.1). Comparisons across groups of interest (*CEBPA* DM, AML with *GATA2* ASE and without) were performed with *methylKit* and average methylation levels were plotted along the *GATA2* gene with *Gviz*[20] (v1.28.3).

Allele-specific methylation of *GATA2* promoters was studied with CRISPR/Cas9-targeted enrichment followed by amplification-free long read sequencing by Oxford Nanopore[21]. Methylation likelihood ratios were estimated with Nanopolish[22] and plotted separately for each allele using *Gviz*.

ChIP-seq and ATAC-seq analyses

ChIP-seq and ATAC-seq data were generated for a number of selected patients to investigate changes in enhancer and promoter regions. ChIP-seq and ATAC-seq were performed as described previously with slight modifications[23,24]. ChIP-seq reads were aligned to the human reference genome build hg19 with *bowtie* and bigwig files were generated for visualization with bedtools genomecov[25] (v2.27.1) and UCSC bedGraphToBigWig[26]. ATAC-seq reads were aligned to the human reference genome build hg19 with *bowtie2*[27] (v2.3.4.1), which is recommended for longer reads, and mitochondrial and duplicate reads were excluded. Bigwig files were generated as described above.

Enhancer regions were defined for quantification of eRNA from RNA-seq, as well as H3K27ac, H3K27me3 and ATAC-seq reads. Read counts in enhancer regions were computed with featureCounts[28] (v1.5.0-p3) and differential analysis was conducted with DESEq2[29] (v1.24.0). The results of this analysis were plotted in the *GATA2* region with *Gviz*[30] (v1.28.3).

An extended description of the methods is provided in the Supplemental Data. Quality metrics for the sequencing data generated in this study are available in Supplementary Table S1.

Data Sharing Statement

Sequence data has been deposited at the European Genome-phenome Archive (EGA, http://www.ebi.ac.uk/ega/), which is hosted by the EBI, under accession number EGAS00001004684.

RESULTS

GATA2 is the most recurrent gene with allele specific expression in AML

To identify instances of epigenetic dysregulation in AML, we performed whole exome sequencing (WES) and RNA-seq on leukemic blasts from 209 AML patient, representing all major subtypes of the disease. Combining both datasets, we assessed ASE in every gene with informative (non-homozygous) single nucleotide variants (SNVs) (Figure 1). Patients had a median of 36 genes with ASE, several of which were recurrently detected across multiple patients (525 in 5 or more patients). The number of genes with ASE was quite stable across patients and was comparable with findings in healthy donors (data not shown), making it unlikely that global mechanisms dictate ASE in AML. No association between genes with ASE in neighboring loci was detected across patients, indicating that causes of ASE were specific to each gene. The degree of ASE, measured by variant allele frequency (VAF) in

the RNA, varied widely across genes and patients: 22% of the ASE events were classified as monoallelic (VAF < 0.1).

To increase the likelihood of disease-relevant observations, we subsequently selected genes previously reported to be involved either in cancer (COSMIC database[31]) or in myeloid development (GO:0030099). Of the genes with ASE complying with these criteria, the 40 most recurrent across the patients of our cohort are included in Table 1 (see Supplementary Table S2 for the complete filtered list). The gene most commonly found to show ASE (37% of cases with informative SNVs) was *GATA2*, which encodes a transcription factor crucial for proliferation and maintenance of hematopoietic stem cells[32].

Table 1. Top 40 genes with recurrent ASE in an AML cohort of 209 patients.

Gene symbol	N	Evaluated cases	% samples	COSMIC	Myeloid Diff
GATA2	66	178	37%	YES	YES
THBS1	36	124	29%	NO	YES
MYH11	20	199	10%	YES	NO
CA2	13	126	10%	NO	YES
MECOM	13	186	7%	YES	NO
SH3PXD2A	13	195	7%	NO	YES
CDKN2A	11	102	11%	YES	NO
JAG1	11	201	5%	NO	YES
L3MBTL3	11	156	7%	NO	YES
TRIM58	11	183	6%	NO	YES
CIB1	10	160	6%	NO	YES
FLT3	10	175	6%	YES	NO
HIP1	9	198	5%	YES	NO
PDE4DIP	9	198	5%	YES	NO
HSP90AB1	8	147	5%	YES	NO
L3MBTL1	8	155	5%	NO	YES
MGMT	8	150	5%	YES	NO
RUNX1	7	188	4%	YES	YES
USP6	7	149	5%	YES	NO
CD101	6	174	3%	NO	YES
FAT1	6	202	3%	YES	NO
IRF8	6	165	4%	NO	YES
MEIS1	6	156	4%	NO	YES
NPM1	6	134	4%	YES	NO
ABL1	5	153	3%	YES	NO
CIITA	5	196	3%	YES	NO

Gene symbol	N	Evaluated cases	% samples	COSMIC	Myeloid Diff
DNMT3A	5	188	3%	YES	NO
FAM20C	5	184	3%	NO	YES
LTF	5	188	3%	NO	YES
MYB	5	161	3%	YES	NO
PML	5	181	3%	YES	YES
PRDM2	5	151	3%	YES	NO
RMI2	5	148	3%	YES	NO
RPN1	5	159	3%	YES	NO
ZFHX3	5	201	2%	YES	NO
AKT1	4	178	2%	YES	NO
BAX	4	110	4%	YES	NO
BRCA1	4	170	2%	YES	NO
KMT2C	4	162	2%	YES	YES
KNSTRN	4	158	3%	YES	NO

The "N" column indicates how many patients present ASE for that gene. "Evaluated cases" indicates how many patients contained SNVs that could be evaluated in that gene. The "% samples" column results from dividing N by the number of evaluated cases. The last two columns indicate whether the gene is found in COSMIC database or is involved in myeloid differentiation (GO:0030099). Note that reportedly imprinted genes (according to GeneImprint) were filtered out.

Molecular lesions in AML exhibit preferential association with gene-specific ASE

Our next question was whether there are preferential associations between genes with ASE and AML-specific mutations. To this end, we selected mutations likely to be somatic (based on their known involvement in AML, presence in COSMIC and pathogenicity predictions) from the variants identified in the WES data (Supplementary Table S3) and calculated the statistical association between every possible pair of mutated genes and genes with ASE (Figure 2).

Unsurprisingly, we found strong associations between driver chromosomal translocations and allele specific expression of their constituent genes: t(11;23) and *KMT2A*, t(8;21) and *RUNX/RUNX1T1*, t(15;17) and *PML*, t(3;3) and *MECOM*, inv(16) and *MYH11*. Upon translocation to a different genomic region, genes previously under the control of another promoter (gene fusions) or enhancer (*MECOM*) acquire monoallelic expression. In addition, the analysis uncovered novel associations between ASE events and mutations, such as *THBS1* with inv(16) (p-value = 0.0008), *MYB* with *ETV6* (p-value = 0.0008), or *LOX* with *SF3B1* (p-value = 0.0028). Among those, the association of *GATA2* ASE with double *CEBPA* mutations (CEBPA DM, p-value = $2.18 \cdot 10^{-5}$) and with *GATA2* mutations (p-value = 0.0004) was the strongest.

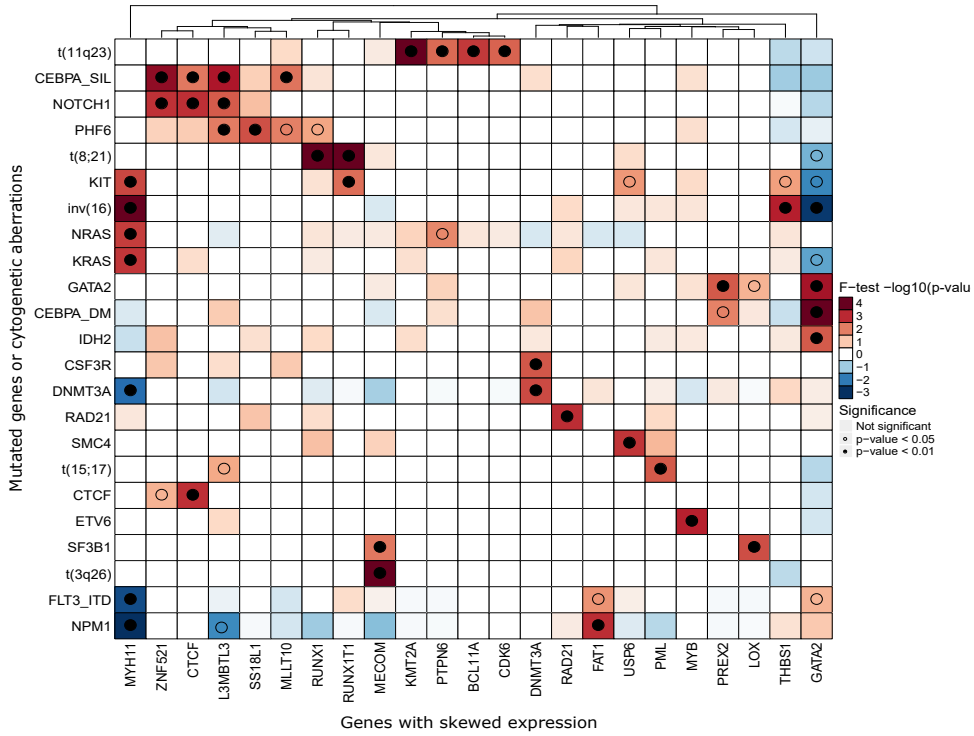


Figure 2. Association between genes with ASE and gene mutations or cytogenetic aberrations. Statistical association was computed with a 2-sided Fisher's exact test and represented as -log10(p-value) for odds ratio > 1 or log10(p-value) for odds ratio < 1. Positive values, indicating positive association, were depicted in red, whereas negative values were depicted in blue. For a clearer visualization, the limits of the scale were set at -4 and +4. Associations that achieve significance were highlighted with an empty (p-value < 0.05) or a full (p-value < 0.01) circle.

GATA2 ASE is strongly associated with *CEBPA* double mutant AML

Given the recurrence of *GATA2* ASE and the prominent role of this gene in leukemogenesis, we further focused on *GATA2*. Therefore, using RNA-seq data, we manually inspected the *GATA2* locus on IGV for all cases to ensure that no case had been excluded by the stringent filtering of our automated pipeline. This second analysis detected *GATA2* ASE in 60% patients with informative SNVs, a substantial increase that was due to the inclusion of UTR regions (absent in the exome-seq data) and the absence of p-value filtering (Supplementary Figure S2). All subsequent calculations are based on this second analysis of the data.

Notably, *GATA2* ASE was detected in all evaluable patients with *CEBPA* double mutations (n=21; p-value = $1.57 \cdot 10^{-5}$, Fisher's test). A statistical analysis of clinically relevant variables revealed other positive associations, albeit weaker, of *GATA2* ASE with normal karyotype, *NPM1* mutations and *FLT3*-ITD mutations. There was no association with white blood cell count, age, sex or ELN 2017 classification (Table 2). Although *GATA2* ASE is widespread in AML, the t(8;21) and t(11q23) subgroups –both involving fusion proteins– were negatively associated with *GATA2* ASE.

Table 2. Clinical characteristics of GATA2 ASE and GATA2 non-ASE groups.

Group	GATA2 ASE (n=103)	GATA2 non-ASE (n=67)	p-value	Effect Size (CI)
Sex				
Female	48% (49)	39% (26)	0.34	0.72 (0.36; 1.4)
Male	49% (50)	55% (37)		
Missing	3.9% (4)	6% (4)		
Age				
Median (MAD)	48.00 (17.79)	47.00 (19.27)	0.79	-0.19 (-0.51; 0.13)
Mean (SD)	48.70% (16.82)	45.57% (16.30)		
Range	15-86	17-77		
Missing	3.9% (4)	6.0% (4)		
ELN classification				
Adverse	20% (21)	30% (20)	0.22	0.14 (0; 0.28)
Favorable	50% (52)	37% (25)		
Intermediate	28% (29)	27% (18)		
Missing	0.97% (1)	6% (4)		
WBC count				
Median (MAD)	43.00 (35.88)	62.00 (52.19)	0.28	0.29 (-0.065; 0.64)
Mean (SD)	60.14% (50.10)	78.29% (80.29)		
Range	1-215	0-510		
Missing	15.5% (16)	26.9% (18)		
NPM1				
Neg	58% (60)	79% (53)	0.005	2.7 (1.3; 6)
Pos	42% (43)	21% (14)		
FLT3-ITD				
Neg	60% (62)	81% (54)	0.0068	2.7 (1.3; 6.2)
Pos	40% (41)	19% (13)		
CEBPA DM				
Neg	80% (82)	100% (67)	<0.001	NA (4; NA)
Pos	20% (21)			
CEBPA SM				
Neg	96% (99)	96% (64)	1	0.86 (0.14; 6.1)
Pos	3.9% (4)	4.5% (3)		
CEBPA silenced				
Neg	94% (97)	93% (62)	0.75	0.77 (0.19; 3.3)
Pos	5.8% (6)	7.5% (5)		
t(15;17)				
Neg	99% (102)	94% (63)	0.079	0.16 (0.0031; 1.6)
Pos	0.97% (1)	6% (4)		
Missing	0% (0)	0% (0)		
t(8;21)				
Neg	99% (102)	93% (62)	0.036	0.12 (0.0026; 1.1)
Pos	0.97% (1)	7.5% (5)		
inv(16)				
Neg	94% (97)	87% (58)	0.1	0.4 (0.11; 1.3)
Pos	5.8% (6)	13% (9)		

Group	GATA2 ASE (n=103)	GATA2 non-ASE (n=67)	p-value	Effect Size (CI)
Normal karyotype				
Neg	36% (37)	67% (45)	<0.001	4.4 (2.1; 9.7)
Pos	57% (59)	24% (16)		
Missing	6.8% (7)	9% (6)		
Complex karyotype				
Neg	70% (72)	64% (43)	0.73	0.75 (0.15; 4)
Pos	4.9% (5)	6% (4)		
Missing	25% (26)	30% (20)		

Descriptive statistics and hypotheses tests were computed for AML patients with or without *GATA* ASE using *Atable*. The p-value and the effect size of statistical tests evaluating association between these groups and clinical variables are shown in the "p-value" and "Effect size" columns, respectively. Effect size has been measured as odds ratio for categorical variables and Cohen's D for numerical variables.
ELN indicates European LeukemiaNet; WBC, white blood cell; CI, confidence interval

GATA2 ASE was not significantly present in other AML subtypes known to be associated with *CEBPA* abnormalities, such as t(8;21)[33] and *CEBPA*-silenced leukemias, both characterized by reduced *CEBPA* expression[34,35] (Figure 3A). Moreover, single *CEBPA* mutations were not associated with *GATA2* ASE either (p-value = 0.708). Therefore, *GATA2* ASE in *CEBPA* DM does not seem to be a general result of abnormalities in *CEBPA* function or expression.

The expressed *GATA2* allele is frequently mutated in AML with CEBPA DM

The second mutated gene with the largest co-occurrence of *GATA2* ASE was *GATA2* itself (p-value = 0.0165). Interestingly, *GATA2* was also mutated in 48% of the *CEBPA* DM cases in our cohort, and 19% carried a second subclonal *GATA2* mutation (Table 3). This is in line with previous findings reporting that 40% of CEBPA DM cases co-occur with *GATA2* mutations[36]. In cases with a *GATA2* mutation, the mutant allele was always preferentially expressed. This suggests a functional connection between *GATA2* and *CEBPA* DM, where ASE may play a cooperative role with *GATA2* mutations.

We did not observe a difference in magnitude of *GATA2* ASE (measured as VAF at RNA level) between *CEBPA* DM patients with or without *GATA2* mutations (Supplementary Figure 3C). Therefore, *GATA2* ASE in *CEBPA* DM occurs independently of the number of *GATA2* mutations.

Our findings were further validated in the TCGA-LAML[37] and in the Beat AML[38] datasets, where all 10 patients with *CEBPA* DM and informative SNVs presented GATA2 ASE (Supplementary Tables S4 and S5). Of these, three patients carried *GATA2* mutations with preferential expression of the mutated allele (Supplementary Figure 4A, B).

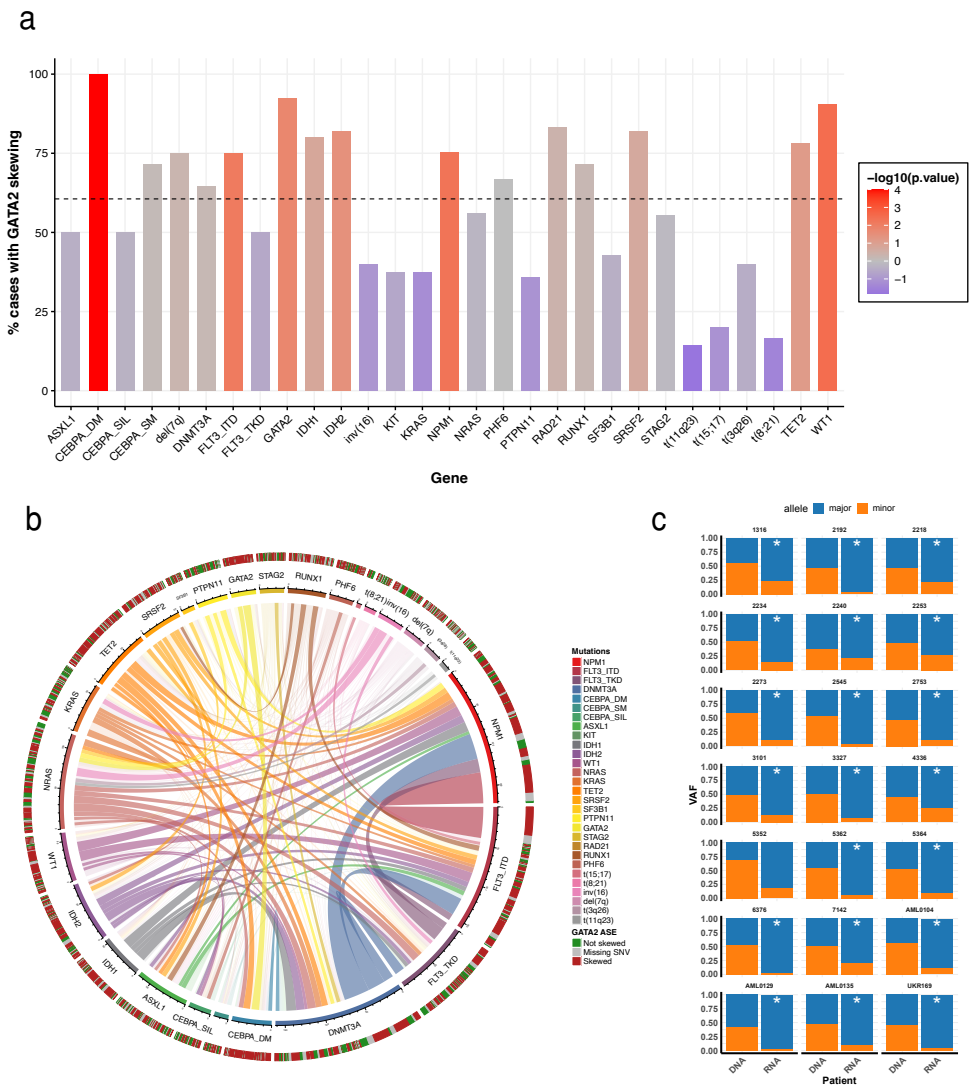


Figure 3. Occurrence of *GATA2* ASE in AML subgroups. (A) Barplot indicating the percentage of cases with *GATA2* ASE in each mutational subgroups. The color of the bars indicates the strength of the association as log10(p-value), with a sign determined by the nature of the association. The scale ranges from blue for negative associations to red for positive associations. The dotted horizontal line indicates the percentage of cases with *GATA2* ASE in the whole AML cohort (B) Circos plot indicating the co-occurrence of mutations in AML and *GATA2* ASE. (C) Bar plots for each *CEBPA* DM patient showing *GATA2* ASE, observed by the discrepancy between VAF at the DNA level and VAF at the RNA level. An asterisk (*) indicates significance at FDR < 0.05 in a chi-square test.

Table 3. *GATA2* and *CEBPA* alterations in *CEBPA* DM patients.

Patient ID	RNA frequency	GATA2 ASE	GATA2 expression	GATA2 mutations		GATA2 allele expressed	CEBPA mutations	CEBPA expression	CEBPA mut VAF	
				N	Type (VAF)				Mut1	Mut2
1316	0.233	SKEWED	106.2	0	.	.	N/C	483.9	0.462	0.448
2192	0.023	MONO	456.2	2	ZF1 (0.39), ZF2 (0.59)	MUT (indel), MUT (0.97)	N/C	390.3	0.526	0.486
2218	0.263	SKEWED	67.8	0	.	.	C/C	308.9	0.923	HMZ
2234	0.144	SKEWED	28.5	2	ZF1 (0.03)	MUT (0.07)	N/C	380.5	0.498	0.475
2240	0.223	SKEWED	41.0	1	ZF1 (0.02)	MUT (0.03)	N/C	328.0	0.486	0.461
2242	.	UNKNOWN	55.5	0	.	.	N/C	162.0	0.472	0.447
2253	0.269	SKEWED	106.2	1	ZF1 (0.47), ZF2 (0.07)	MUT (0.71), MUT (0.49)	N/C	168.1	0.490	0.418
2273	0.0993	MONO	61.0	1	ZF1 (0.47)	MUT (0.92)	N/C	161.4	0.488	0.423
2545	0.037	MONO	106.5	1	ZF2 (0.39)	MUT (0.96)	N/C	274.7	0.497	0.484
2753	0.106	SKEWED	40.9	1	ZF1 (0.45)	MUT (0.93)	N/C	233.7	0.448	0.441
3101*	0.126	SKEWED	50.9	0	.	.	N/N	194.4	NA	NA
3327	0.071	MONO	94.1	0	.	.	C/C	86.2	0.918	HMZ
4336	0.285	SKEWED	36.7	0	.	.	N/C	143.7	0.442	0.470
5352	0.174	SKEWED	24.3	0	.	.	N/C	417.6	0.472	0.412
5362	0.064	MONO	60.2	2	ZF1 (0.03), ZF2 (0.49)	MUT (0.12), MUT (0.93)	N/C	238.8	0.497	0.464
5364	0.097	MONO	113.9	0	.	.	N/N	427.4	0.283	0.277
6376	0.024	MONO	43.4	0	.	.	C/C	258.7	0.899	HMZ
7142	0.208	SKEWED	29.7	0	.	.	N/C	141.2	0.482	0.473
AML0104	0.107	MONO	66.6	0	.	.	C/C	264.1	0.422	HMZ
AML0129†	0.018	MONO	10.1	0	.	.	N/N	169.5	0.035	0.334
AML0135	0.097	MONO	60.3	2	ZF1 (0.19), ZF2 (0.37)	MUT (0.46), MUT (0.87)	N/C	125.0	0.399	0.173
UKR169	0.051	MONO	13.9	1	ZF1 (0.45)	MUT (0.96)	N/C	318.8	0.847	HMZ

"RNA frequency" indicates the proportion of reads that come from the minor allele for all the SNPs considered in the gene. "GATA2 ASE" was categorized as "monoallelic" for RNA frequency <= 0.10 or "skewed" for RNA frequency <= 0.35. The expression of *GATA2* and *CEBPA* is presented in TPM as reported by Salmon. The "GATA2 mutations" column contains the number, type (ZF1/2) and VAF of the mutations identified in *GATA2*. The "GATA2 allele expressed" column includes the VAF of these same mutations measured in the RNA. The VAF of the two *CEBPA* mutations, based on deep amplicon-sequencing, is indicated in N- to C-terminus order.

MONO indicates monoallelic expression; VAF, variant allele frequency; ZF, zinc finger; MUT, mutated allele; N/C, N-terminal/C-terminal mutation; HMZ, homozygous

*Amplicon-sequencing was not conducted for 3101 and *CEBPA* VAF is unavailable.

†AML0129 has a *CEBPA* mutation in only one allele, but the other allele is not expressed; thus, it acts like a *CEBPA* homozygous mutation at the transcriptional level.

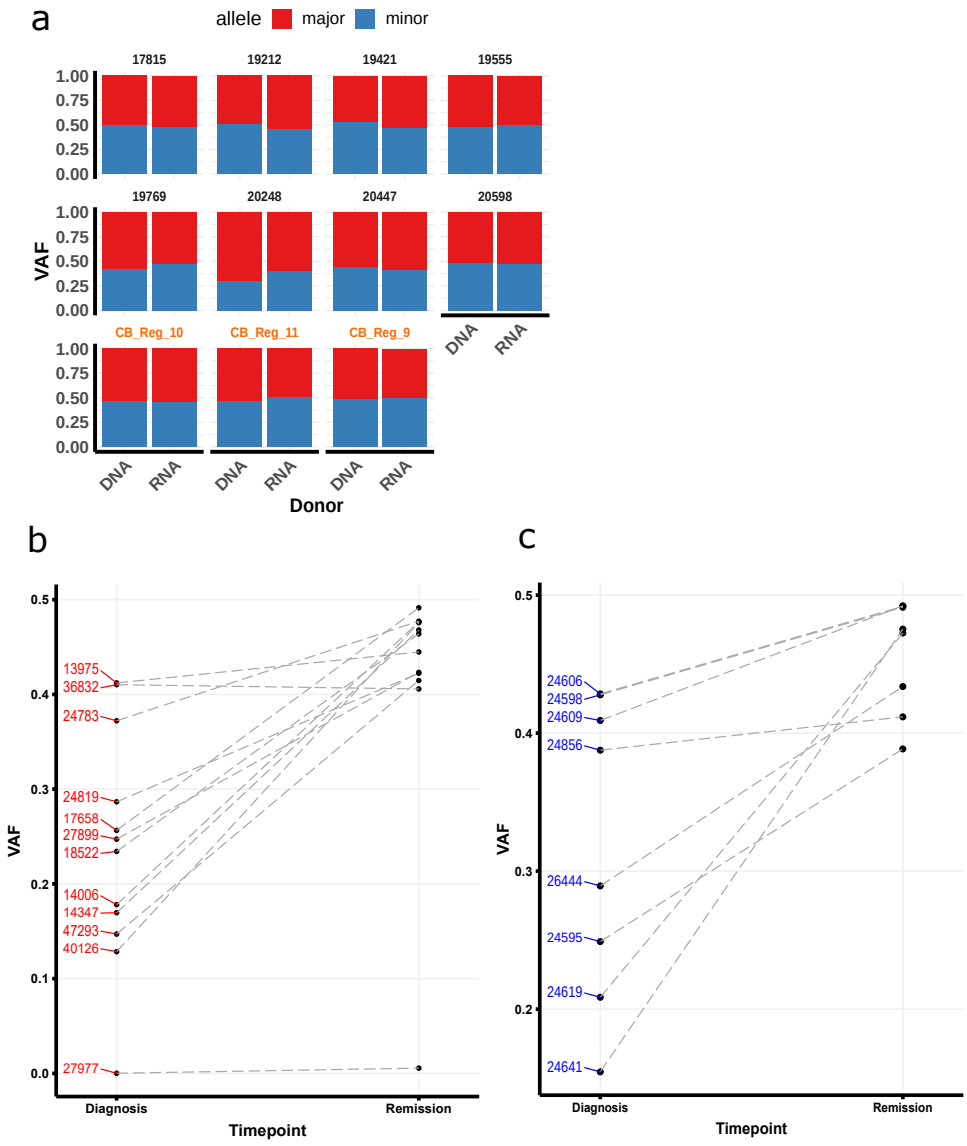


Figure 4. *GATA2* ASE is only present in leukemia cells. (A) Bar plot showing the absence of *GATA2* ASE in CD34+ cells, of which 8 derived from bone marrow and 3 from cord blood (in orange). The average variant allele frequency (VAF) along the *GATA2* gene at the DNA and the RNA levels is identical in all samples. (B) Comparison of VAF measured in RNA at diagnosis or remission in *CEBPA* DM samples. (C) Comparison of VAF measured in RNA at diagnosis or remission in *NPM1* mutated samples.

GATA2 ASE is a somatic event in *CEBPA* DM AML

Our observations suggest a role of *GATA2* ASE in the pathogenesis of *CEBPA* DM AML, which would imply that *GATA2* ASE should be leukemia-specific and not present in healthy controls. An analysis of bone marrow- (n=8) or cord blood-derived (n=3) hematopoietic stem cells from healthy individuals did not show any *GATA2* ASE, indicating that *GATA2* ASE is not commonly found in the general population (Figure 4A).

To examine whether *GATA2* ASE is indeed present at the time of leukemia development and lost upon achieving remission following treatment, we sequenced a second series of *CEBPA* DM cases (n=12) for which both diagnostic and complete remission material was available (Table 4). In these cases targeted *GATA2* DNA and cDNA amplicon sequencing was applied, having previously confirmed that this technique recapitulates the RNA-seq results (Supplementary Figure 5). In the diagnostic samples, we again observed frequent *GATA2* ASE, although slightly less frequent than in the previous series (10/12 cases, 83%).

At remission, biallelic expression of *GATA2* was restored in 9 out of 10 *CEBPA* DM samples which showed *GATA2* ASE at diagnosis (Figure 4B, Supplementary Figure S6A). The exception, case 27977, displayed completely monoallelic expression of *GATA2* at both timepoints, potentially indicating that *GATA2* ASE preceded leukemia development in that particular patient. Interestingly, that same patient exhibited one N-terminal *CEBPA* mutation in 50% of the cells in remission, suggesting that it carried a germline *CEBPA* mutation accompanied by germline *GATA2* ASE. In a control group of AML with *NPM1* mutations with *GATA2* ASE at diagnosis we similarly observed *GATA2* biallelic expression at remission (Figure 4C, Supplementary Figure S6B).

Overall, these data indicate that *GATA2* ASE is a leukemia-specific event since it is absent in healthy cells and is lost in complete remission.

GATA2 promoters are differentially methylated in CEBPA DM AML

Methylation of CpG islands proximal to a transcriptional start site (TSS) may block transcription initiation and is correlated with loss of gene expression[39]. To explore this in the context of *GATA2* ASE, we analyzed enhanced reduced representation bisulfite sequencing (ERRBS) data generated in a subset (n=35) of our AML cohort[18].

The *GATA2* gene encodes multiple isoforms with different TSS, all of which overlap with a long CpG island. We defined promoters as the 1000 bp regions upstream of the TSS of isoforms expressed in AML: a short (Prom-S) and a long (Prom-L) isoform (Supplementary Figure S7A). We compared methylation levels in these promoters for the following three groups (Figure 5A, Supplementary Figure S7B): 1. *CEBPA* DM AML with *GATA2* ASE (CEBPA_DM, n=10), 2. AML without *CEBPA* DM but with GATA2 ASE (Control_ASE, n=20) and AML without *CEBPA* DM and without *GATA2* ASE (Control_BE, n=5). We identified significant hypermethylation in *CEBPA* DM in the promoter of the long *GATA2* form with respect to Control_ASE (p-value < 0.0001), but not Control_BE (p-value = 0.0016). No significant differences were observed in the promoter of the short isoform.

For further validation, we conducted bisulfite treatment followed by amplicon-seq of *GATA2* promoters in additional samples from the original cohort: 9 CEBPA_DM, 7 Control_ASE and 2 Control_BE. Here, the regions were more narrowly defined, but sequenced with a higher resolution than that achieved by ERRBS. The results confirmed the previous observations (Figure 5B, Supplementary Figure S8A): the CEBPA_DM group exhibited hypermethylation in the promoter of the long *GATA2* form when compared to Control_ASE (p-value < 0.0001) and Control_BE (p-value = 0.0571). Moreover, we conducted bisulfite sequencing on 4 paired diagnosis-remission samples of *CEBPA* DM where we had previously detected *GATA2* ASE (Figure 4). In all cases, we observed a strong decline of methylation levels in Prom-L at remission, consistent with the notion that hypermethylation associated with *GATA2* ASE is a leukemia-specific event (Figure 5C, Supplementary Figure S8B).

Table 4. Characteristics of *CEBPA* double mutant patients with remission material available.

Patient ID	RNA freq D	Skewing	RNA freq R	GATA2 mutations		GATA2 allele expressed	CEBPA mutations	CEBPA mutations VAF			
				Num.	Type (VAF)			Mut1-D	Mut2-D	Mut1-R	Mut2-R
13975	41.21%	NOT SKEWED	44.48%	0		.	N/C	0.396	0.459	0.000	0.000
14006	17.81%	SKEWED	47.66%	0		.	N/N	0.882	HMZ	0.000	HMZ
14347	16.96%	SKEWED	46.79%	1	ZF1 (0.49)	MUT (0.82)	N/C	0.457	0.420	0.000	0.183
17658	25.65%	SKEWED	49.15%	0		.	N/N	0.457	0.460	0.001	0.000
18522	23.43%	SKEWED	46.39%	0		.	C/C	0.781	HMZ	0.001	HMZ
24783	37.22%	SKEWED	47.73%	0		.	N/C	0.446	0.436	0.000	0.000
24819	28.66%	SKEWED	42.23%	1	ZF1 (0.06)	MUT (0.10)	N/C	0.401	0.316	0.000	0.000
27899	24.73%	SKEWED	42.33%	0		.	N/C	0.470	0.460	0.000	0.000
27977	0.01%	MONO	0.56%	0		.	N/C	0.503	0.434	0.000	0.501
36832	41.03%	NOT SKEWED	40.58%	0		.	N/C	0.438	0.389	0.000	0.000
40126	12.85%	SKEWED	47.57%	2	ZF1 (0.12), ZF1 (0.07)	MUT (0.16), MUT (0.11)	N/C	0.469	0.469	0.000	0.000
47293	14.71%	SKEWED	41.46%	0		.	N/C	0.435	0.459	0.000	0.001

"RNA frequency" indicates the proportion of reads that come from the minor allele for each SNP considered, and has been determined at diagnosis (D) or remission (R). "GATA2 skewing" was categorized as "monoallelic" (MONO) for RNA frequency <= 0.10 or "skewed" for RNA frequency <= 0.35. The expression of *GATA2* and *CEBPA* is presented in TPM as reported by Salmon. The "*GATA2* mutations" column reports the VAF of the mutation at the DNA level between parentheses, whereas the "GATA2 allele expressed" column includes the VAF measured in the RNA. The VAF of the two CEBPA mutations is indicated in N- to C-terminus order, at diagnosis (D) and remission (R). MONO indicates monoallelic expression; VAF, variant allele frequency; ZF, zinc finger; MUT, mutated allele; N/C, N-terminal/C-terminal mutation; HMZ, homozygous

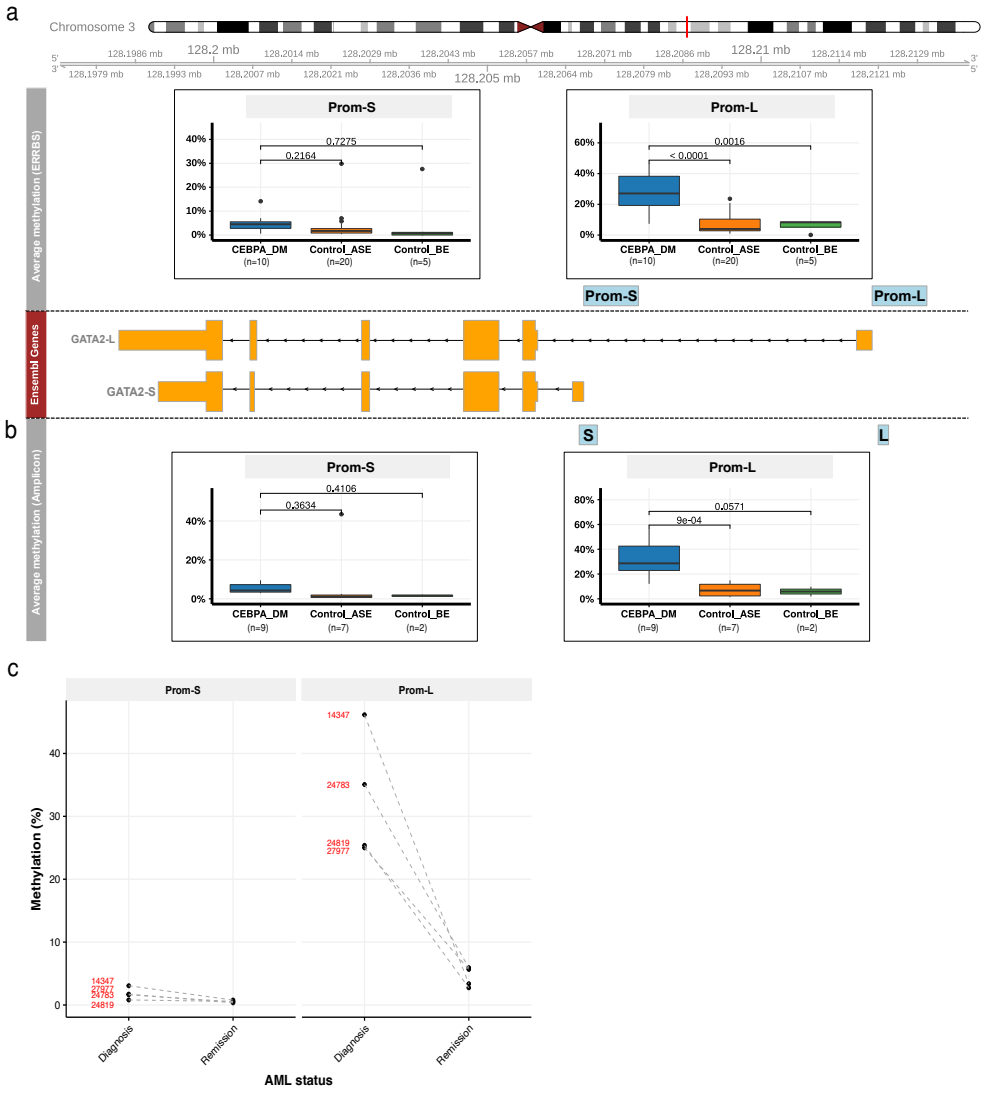


Figure 5. Methylation analysis of *GATA2* promoters. (A) Differential methylation analysis of putative promoters of the two expressed *GATA2* isoforms using ERRBS (Prom-S and Prom-L for Short and Long respectively). The following groups were compared: AMLs with *CEBPA* DM (CEBPA_DM, n=10), other AMLs with *GATA2* ASE (Control_ASE, n=20) and AMLs with biallelic *GATA2* expression (Control_BE, n=5). The y axis indicates the percentage (%) of methylation, averaged for all the CpG positions in each promoter region. (B) Differential methylation analysis of the promoters of the two expressed *GATA2* isoforms using bisulfite treatment followed by amplicon-sequencing. Note that the amplified regions (denoted as S and L) are selections of the sequences examined in the ERRBS data. Groups were defined as described: CEBPA_DM (n=9), Control_ASE (n=7), Control_BE (n=2). (C) Methylation changes in *GATA2* promoters of paired diagnosis-remission samples from *CEBPA* DM AML patients.

Methylation of *GATA2* promoters is allele-specific and correlates with expression

To confirm that the less transcriptionally active *GATA2* allele is repressed via methylation, we carried out CRISPR/Cas9-targeted enrichment of the *GATA2* locus followed by amplification-free long read sequencing in 4 *CEBPA* DM patients by Oxford Nanopore, which allows direct detection of methylation[40]. We estimated CpG methylation likelihood in each allele separately, based on a heterozygous SNP that also enabled ASE detection.

In general, the individual methylation patterns recapitulated the ERRBS data (Supplementary Figure S9A). The results were also consistent across different methylation callers (Supplementary Figure S9B). Interestingly, there were no differences in Prom-L between the two alleles, both of which were strongly methylated (Figure 6, Supplementary Figure S9C). Although CEBPA DM patients are uniquely methylated in this region, certain positions exhibited 100% methylation in the selected patients (Supplementary Figure S8A). This is incompatible with allele-specific methylation, and thus in line with the Nanopore results. On the other hand, 3/4 patients presented allelic specific methylation of the less abundant allele in Prom-S. This further supports the notion that the less transcriptionally active *GATA2* allele is repressed via methylation in *CEBPA* DM.

GATA2 levels appear to be preserved by a compensatory mechanism involving its -110 kb enhancer

Comparing expression levels across the abovementioned groups, there was no loss of *GATA2* transcript levels in AML patients with *CEBPA* DM (Figure 7A). We hypothesized that changes in the activity of a *GATA2* enhancer in *cis* may compensate the absence of transcription from the other allele. The promoters of *GATA2* interact with a variety of *cis*-regulatory elements that dictate tissue-specific expression, including the 9.5 kb intronic enhancer and the -110 kb distant super-enhancer[41]. The -110 kb enhancer (-77 kb in mice) is essential for embryogenesis, controls differentiation of CMPs and GMPs[42], and its loss is involved in development of AML with inv(3)/t(3;3)[4]. Therefore, we examined changes in the activity of this enhancer.

Differential expression analysis revealed that *CEBPA* DM cases exhibit increased transcription in all the elements contained within the *GATA2* super-enhancer (p-value < 0.05, *DESeq2*) when compared to other AML cases, regardless of whether they exhibit *GATA2* ASE or not (Figure 7B). Increased transcription in enhancer regions was shown to be allele-specific for 4/6 CEBPA DM samples where DNA sequencing information was available in that region (Figure 7C). Likewise, levels of both H3K27ac (Figure 7D) and ATAC-seq (Supplementary Figure S10A) were higher for CEBPA DM cases than any other group in the GATA2 super-enhancer region. Interestingly, the patterns of allele specificity sometimes differed between enhancer RNA (eRNA) and H3K27ac data (Figure 7C, Figure 7E).

There were no significant differences in super-enhancer methylation, although it should be noted that the resolution of ERRBS in this area was low (Supplementary Figure S10B). Besides, there were no differences in H3K27me3 (Supplementary Figure S10C), a mark for

poised enhancers[43]. H3K27me3, which is mediated by the polycomb complex PRC2, is also present in the promoters of silenced genes and might prevent transcription[44]. However, we did not observe significant differences in any of the *GATA2* locus regions examined, ruling out PRC2-mediated repression (Supplementary Figure S10D).

Altogether, these results support the notion that inactivation of one *GATA2* allele by methylation is compensated by increased enhancer activity in the other allele, leading to maintenance of *GATA2* levels.

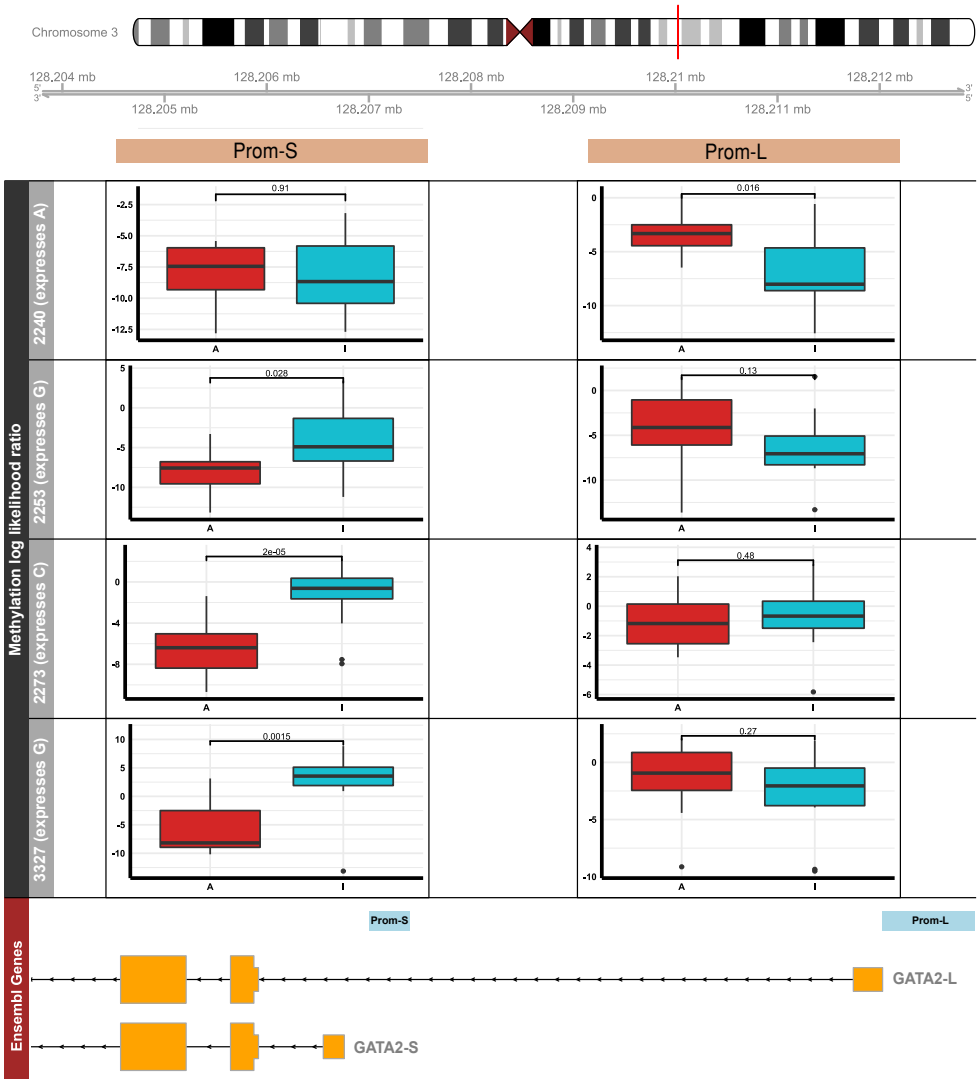


Figure 6. Detection of allele-specific methylation in *GATA2* promoters. Differential methylation analysis of putative promoters of the two expressed *GATA2* isoforms by Nanopore sequencing (Prom-S and Prom-L for Short and Long respectively). In 4 *CEBPA* DM patients, the more abundant allele (A) was compared to the less transcriptionally active allele (I) based on a heterozygous SNP: rs72983369 for 2240 and rs1573858 for 2253, 2273 and 3327. Methylation likelihood ratios computed by Nanopolish were averaged across all reads mapping to each allele.

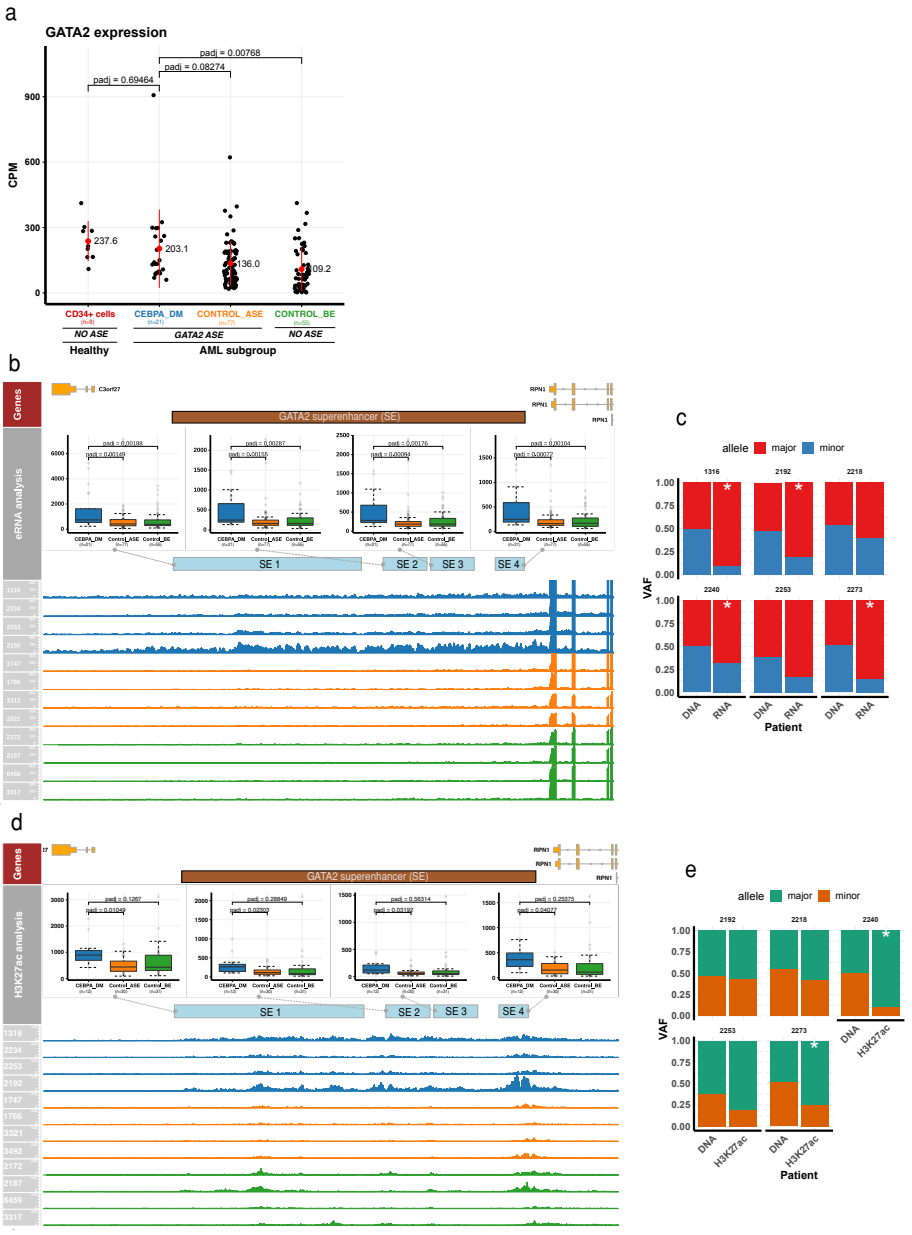


Figure 7. Compensation of *GATA2* levels by super-enhancer activation. (A) Comparison of *GATA2* expression levels in AML groups and CD34+ normal control cells (n=9). The following AML groups were compared: AMLs with *CEBPA* DM (CEBPA_DM, n=21), other AMLs with *GATA2* ASE (Control_ASE, n=77) and AMLs with biallelic *GATA2* expression (Control_BE, n=55). No loss of *GATA2* expression is observed in *CEBPA* DM. (B) Analysis of eRNA expression in the *GATA2* -110 kb super-enhancer. (C) Allele specific expression of eRNA in the *GATA2* super-enhancer, comparing CEBPA_DM (n=21), Control_ASE (n=77) and Control_BE (n=55). The variant allele frequency (VAF) of the DNA and the eRNA are shown. An asterisk (*) indicates significance at FDR < 0.05 in a chi-square test. (D) Analysis of H3K27ac binding levels in the *GATA2* -110kb super-enhancer, comparing CEBPA_DM (n=12), Control_ASE (n=30) and Control_BE (n=31) (E) Allele specific binding of H3K27ac in the *GATA2* super-enhancer. The variant allele frequency (VAF) of the DNA and the H3K27ac reads are shown. An asterisk (*) indicates significance at FDR < 0.05 in a chi-square test

DISCUSSION

We detected *GATA2* ASE in 60% of the AML cases, with a very strong association with *CEBPA* DM. Analysis of additional cohorts revealed that *GATA2* ASE is found in 41/43 (95%) of *CEBPA* DM AML cases, and is a somatic, leukemia-specific event that is lost upon remission. In cases with *GATA2* mutations, the mutated allele was preferentially expressed, but ASE was also present in the absence of *GATA2* mutations. We show that our findings can be explained by simultaneous silencing of one allele by methylation and overactivation of the other allele via the -110 kb super-enhancer, resulting in unchanged, or even slightly increased, *GATA2* levels. Collectively, these data suggest that *GATA2* ASE is an important event in the development of AML with *CEBPA* DM.

GATA2 encodes a transcription factor crucial for proliferation and maintenance of hematopoietic stem cells[32]. Balanced expression of functional *GATA2* is critical for normal hematopoiesis, with alterations in either its expression or activity having been linked to leukemogenesis[45]. For instance, gain-of-function *GATA2* mutations mediate acute myeloid transformation of chronic myeloid leukemia[46], whereas loss-of-function germline mutations leading to *GATA2* deficiency predispose carriers to familial myelodysplastic syndrome (MDS)/AML[47]. These patients present a wide range of other phenotypic manifestations including immunodeficiency, pulmonary disease and lymphatic dysfunction[48]. Besides mutations in coding regions of the gene, these symptoms can be caused by mutations in an internal enhancer of *GATA2,* resulting in reduced expression of the gene product[49]. On the other hand, *GATA2* overexpression has been suggested to be a poor prognostic marker in both pediatric[50] and adult[51] AML. Our findings demonstrate that *GATA2* defects may not only be caused by mutations in the gene or its regulatory elements, but underscores the importance of epigenetic changes or “epimutations” in this gene in a subset of leukemias.

These observations highlight the importance of a fine-tuned regulation of *GATA2* expression and point to a role of *GATA2* ASE in the pathogenesis of AML. Accordingly, Celton et al.[52] also reported frequent *GATA2* ASE in a smaller cohort of 49 normal karyotype AML patients, though it should be noted that other genes were not considered in that study. In a much larger group of patients, we conclusively demonstrate that *GATA2* displays ASE more often than any other known myeloid or cancer-related gene. Moreover, although *GATA2* ASE is widespread in AML, we show it is distinctly associated with *CEBPA* DM, where both events co-occur in 95% of the 43 cases analyzed.

Double *CEBPA* mutations define an AML subtype with a distinct gene expression profile and favorable clinical outcome[53,54]. These patients typically exhibit a combination of an N- and C-terminal mutations in the CEBPA protein, disrupting its dimerization and DNA-binding activities[55]. We did not find an association between *GATA2* ASE and the type of *CEBPA* mutations present in each patient (Supplementary Figure 3D).

The specific association between *GATA2 ASE* and *CEBPA* DM suggests cooperativity of these two genes in the context of leukemogenesis. This is in keeping with the previously reported observation that *GATA2* mutations are present in approximately 40% of the *CEBPA* DM cases. Somatic *GATA2* mutations mainly cluster in the two zinc finger (ZF) domains of the protein, each with different functional implications[56]. The ZF1 domain (N-terminal) of GATA2 contributes to the stabilization and specificity of DNA binding and mediates the interaction with FOG1, whereas the ZF2 interacts with CEBPA[36]. The role of these mutations in AML is a subject of ongoing research, with effects described on proliferation and differentiation (see Leubolt et al.[56] for a recent review). ZF1 mutations are strongly associated with *CEBPA* DM, where they may play a cooperative role: the mutations lead to reduced transcription of CEBPA targets[36]. All the cases of our cohort with *GATA2* mutations exhibited at least an aminoacid change in ZF1, but those with two mutations had a second hit in ZF2. Strikingly, both *GATA2* mutations were always in the same allele, which was preferentially expressed. In a recent study of recurrently mutated genes in AML, Batcha et al.[57] also identified allelic imbalance towards mutant *GATA2*, though their effort was limited to 11 genes harboring recurring mutations. Similarly, Al Seraihi et al. reported *GATA2* ASE favoring the mutated allele in a family with inherited *GATA2*-mutated MDS/AML[58]. In contrast, Kozyra et al. recently described synonymous *GATA2* mutations in patients with MDS that lead to decreased transcript stability, leading to ASE favoring the wild type alele[59]. In *CEBPA* DM AML patients with *GATA2* mutations, the presence of *GATA2* ASE can be explained because it leads to dominance of the mutated allele. However, because *GATA2* ASE was also observed in the vast majority of *CEBPA* DM cases without *GATA2* mutations, we hypothesize that *GATA2* ASE precedes the acquisition of mutations.

The average expression of *GATA2* in *CEBPA* DM patients was comparable to other AMLs, even in cases with monoallelic *GATA2* expression. We show that this is due to DNA methylation-mediated gene silencing of the repressed allele, compensated by overactivation of the long-distance -110 kb GATA2 super-enhancer on the other allele (Supplementary Figure S11). Interestingly, this is the same regulatory element involved in AML with t(3;3)/inv(3)[4], as well as many other atypical 3q26 translocations[60]. However, in these leukemias loss of the *GATA2* super-enhancer results in GATA2 haploinsufficiency, which accelerates *EVI1*-driven leukemogenesis[61]. Given the very strong association between GATA2 ASE and *CEBPA* DM, we hypothesize that *GATA2* ASE also contributes to *CEBPA*-mediated leukemogenesis, albeit the exact mechanisms remain unclear. One possibility is that silencing of one allele and enhancer activation of the other allele do not originate at the same time. Instead, high levels of *GATA2* driven by the -100 kb enhancer may contribute to leukemia initiation in preleukemic cells, whereas loss of expression may be favored in later stages. This hypothesis is consistent with the findings by Saida et al. in inv(16) AML models, where *Gata2* expression is upregulated in the preleukemic phase, but monoallelic *Gata2* deletions lead to a more aggressive phenotype in the leukemic stage[62]. Studies using *Cebpa* DM mouse leukemias *in vivo*[63] could further clarify the order of acquisition of *Gata2* ASE in those leukemias.

The acquisition of methylation and acetylation marks in the absence of changes in the DNA constitutes an example of "epimutation"[64]. Such epigenetic modifications have been extensively detected in cancer, often affecting the expression levels of tumor suppressor genes[65]. Here, we show that epimutations leading to *GATA2* ASE are mostly somatic and lost at remission, which further supports the notion that they play a role in leukemia development. Although hyperactivation of the -110 kb super-enhancer was not reported, other studies had previously detected hypermethylation of the *GATA2* promoter in non-*CEBPA* DM cases[52,58]. Why *GATA2* is prone to acquisition of these epimutations and how or when they are exactly incorporated remains to be elucidated. One intriguing possibility is that *GATA2* ASE is acquired at a certain differentiation stage that becomes the leukemia cell of origin. Given that other subgroups with *CEBPA* abnormalities (other than mutations) do not show a similar pattern, we propose that ASE of *GATA2* is not a consequence of *CEBPA* mutations. Intriguingly, methylation levels of other AML cases with *GATA2* ASE are low, suggesting there might be another mechanism at play in those.

In a single patient with *CEBPA* DM we observed *GATA2* ASE at diagnosis as well as in remission, which poses several questions for future research. First, *GATA2* ASE in remission marrow should be analyzed in a much larger cohort to determine the frequency of such a condition. Second, it would be interesting to determine whether *GATA2* ASE was already present in bone marrow progenitors before leukemic transformation and, if so, whether it was somatically acquired or present in the germline. Importantly, this would suggest that an SNV in a regulatory domain of *GATA2* is responsible for such an effect.

In summary, *GATA2* ASE is a somatic event that is epigenetically acquired in almost all *CEBPA* DM AML cases, suggesting it plays a key role in the development and/or progression of this leukemia subtype – a notion further supported by the association between *GATA2* mutations and *CEBPA* mutations. The specific mechanisms remain unclear, but the importance of fine-tuned *GATA2* regulation points to *GATA2* levels. Thus, we propose that increased levels of *GATA2* mediated by over-activation of the super-enhancer, in collaboration with *CEBPA* mutations, might be an early event in leukemic transformation. Later, allele-specific silencing would result in stabilization of *GATA2* levels in leukemic blasts.

AUTHOR CONTRIBUTIONS

Contribution: R.M., B.J.W. and R.D designed the study; S.v.H., C.E., C.G. and I.R. carried out experiments; R.M., M.A.S., C.V., J.d.R. and P.J.M.V. analyzed data; P.J.M.V., A.M.M. and M.R. provided samples and/or data; R.M., B.J.W. and R.D. wrote the manuscript.
Correspondence: Bas Wouters, Department of Hematology and Oncode Institute, Erasmus University Medical Center, Wijtemaweg 80, 3015CN Rotterdam (The Netherlands); email: b.wouters@erasmusmc.nl
My contributions to this work were: design of the study; implementation of the pipeline for the detection of allele-specific expression; processing and analysis of all high throughput sequencing data (WES, RNA-seq, ERRBS, ChIP-seq, ATAC-seq); data management and upload; interpretation of the results and writing of the manuscript.

ACKNOWLEDGEMENTS

The authors are indebted to the colleagues from the bone marrow transplantation group and the molecular diagnostics laboratory of the department of Hematology at the Erasmus University Medical Center for storage of samples and molecular analysis of the leukemia cells. We also thank our colleagues of the Hematology department for their input, and especially Remco Hoogenboezem for bioinformatics support and algorithm implementation. Moreover, we would like to mention research technicians involved in this work: Margit Nützel, Hanna Stanewsky, Johanna Raithel and Ute Ackermann. We are also grateful to Roberto Avellino for critically reading the manuscript and to Timothy Ley for discussing the findings. This work was funded by grants and fellowships from the Dutch Cancer Society, “Koningin Wilhelmina Fonds” (R. D., B.J.W, R.M., S.v.H.) and a Special Fellowship Award by the Leukemia & Lymphoma Society (B.J.W.). A.M.M. is funded by NCI UG1 CA233332, NCI R01 CA198089 and LLS SCOR 7013-17.

CONFLICT OF INTEREST DISCLOSURE

Conflict-of-interest disclosure: A.M.M. receives research funding from Janssen, Daiichi Sankyo and Sanofi. He has consulted for Epizyme, Constellation, BMI and Exo-Therapeutics. He is a scientific advisor to KDAC. J.d.R is co-founder of Cyclomics BV. The rest of the authors declare no competing financial interests.

REFERENCES

1. Bradner JE, Hnisz D, Young RA. Transcriptional Addiction in Cancer. *Cell*. 2017;168(4):629–643.
2. Papaemmanuil E, Gerstung M, Bullinger L, et al. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N. Engl. J. Med.* 2016;374(23):2209–2221.
3. Bhagwat AS, Lu B, Vakoc CR. Enhancer dysfunction in leukemia. *Blood*. 2018;131(16):1795–1804.
4. Gröschel S, Sanders MA, Hoogenboezem R, et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell*. 2014;157(2):369–381.
5. Shi J, Whyte WA, Zepeda-Mendoza CJ, et al. Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation. *Genes Dev.* 2013;27(24):2648–2662.
6. Mansour MR, Abraham BJ, Anders L, et al. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science (80-.).* 2014;346(6215):1373–1377.
7. Guo YA, Chang MM, Huang W, et al. Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. *Nat. Commun.* 2018;9(1):.
8. Kulis M, Esteller M. DNA Methylation and Cancer. Adv Genet; 2010.
9. Flavahan WA, Drier Y, Liau BB, et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*. 2016;529(7584):110–114.
10. Sur I, Taipale J. The role of enhancers in cancer. *Nat. Rev. Cancer*. 2016;16(8):483–493.
11. Pastinen T. Genome-wide allele-specific analysis: Insights into regulatory variation. *Nat. Rev. Genet.* 2010;11(8):533–538.
12. Clayton EA, Khalid S, Ban D, et al. Tumor suppressor genes and allele-specific expression: mechanisms and significance. *Oncotarget*. 2020;11(4):462–479.
13. Valle L, Serena-Acedo T, Liyanarachchi S, et al. Germline allele-specific expression of TGFBR1 confers an increased risk of colorectal cancer. *Science (80-.).* 2008;321(5894):1361–1365.
14. Van Driel WJ, Tjiong MY, Hilders CGJM, Trimbos BJ, Fleuren GJ. Association of allele-specific HLA expression and histopathologic progression of cervical carcinoma. *Gynecol. Oncol.* 1996;62(1):33–41.
15. Liu Z, Dong X, Li Y. A Genome-Wide Study of Allele-Specific Expression in Colorectal Cancer. *Front. Genet.* 2018;9:570.
16. Lee S, Kim J, Lee S. A comparative study on gene-set analysis methods for assessing differential expression associated with the survival phenotype. *BMC Bioinformatics*. 2011;12:377.
17. Ströbel A. atable: Create Tables for Clinical Trial Reports. *R J.* 2019;11(1):137–148.
18. Glass JL, Hassane D, Wouters BJ, et al. Epigenetic identity in AML depends on disruption of nonpromoter regulatory elements and is affected by antagonistic effects of mutations in epigenetic modifiers. *Cancer Discov.* 2017;7(8):868–883.
19. Akalin A, Kormaksson M, Li S, et al. MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* 2012;13(10):1–9.
20. Hahne F, Ivanek R. Visualizing genomic data using Gviz and bioconductor. *Methods Mol. Biol.* 2016;1418:335–351.
21. Stangl C, de Blank S, Renkens I, et al. Partner independent fusion gene detection by multiplexed CRISPR-Cas9 enrichment and long read nanopore sequencing. *Nat. Commun.* 2020;11(1):1–14.
22. Simpson JT, Workman RE, Zuzarte PC, et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods*. 2017;14(4):407–410.
23. Pham TH, Benner C, Lichtinger M, et al. Dynamic epigenetic enhancer signatures reveal key transcription factors associated with monocytic differentiation states. *Blood*. 2012;119(24):e161–e171.

24. Corces MR, Trevino AE, Hamilton EG, et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods*. 2017;14(10):959–962.
25. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–842.
26. Speir ML, Zweig AS, Rosenbloom KR, et al. The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.* 2016;44(D1):D717-25.
27. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. 2012;9(4):357–359.
28. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923–930.
29. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550.
30. Hahne F, Ivanek R, Lalonde E, et al. Visualizing Genomic Data Using Gviz and Bioconductor. *Source Code Biol. Med. 2016 111*. 2016;1418(43):335–351.
31. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* 2019;47(D1):D941–D947.
32. Tsai F-Y, Orkin SH. Transcription Factor GATA-2 Is Required for Proliferation/Survival of Early Hematopoietic Cells and Mast Cell Formation, But Not for Erythroid and Myeloid Terminal Differentiation. *Blood*. 1997;89(10):3636–3643.
33. Pabst T, Mueller BU, Harakawa N, et al. AML1-ETO downregulates the granulocytic differentiation factor C/EBPα in t(8;21) myeloid leukemia. *Nat. Med.* 2001;7(4):444–451.
34. Wouters BJ, Jordà MA, Keeshan K, et al. Distinct gene expression profiles of acute myeloid/T-lymphoid leukemia with silenced CEBPA and mutations in NOTCH1. *Blood*. 2007;110(10):3706–3714.
35. Figueroa ME, Wouters BJ, Skrabanek L, et al. Genome-wide epigenetic analysis delineates a biologically distinct immature acute leukemia with myeloid/T-lymphoid features. *Blood*. 2009;113(12):2795–804.
36. Greif PA, Dufour A, Konstandin NP, et al. GATA2 zinc finger 1 mutations associated with biallelic CEBPA mutations define a unique genetic entity of acute myeloid leukemia. *Blood*. 2012;120(2):395–403.
37. Ley TJ. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *N. Engl. J. Med.* 2013;368(22):2059–2074.
38. Tyner JW, Tognon CE, Bottomly D, et al. Functional genomic landscape of acute myeloid leukaemia. *Nature*. 2018;562(7728):526–531.
39. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* 2012;13(7):484–492.
40. Wescoe ZL, Schreiber J, Akeson M. Nanopores discriminate among five C5-cytosine variants in DNA. *J. Am. Chem. Soc.* 2014;136(47):16582–16587.
41. Wlodarski MW, Collin M, Horwitz MS. GATA2 deficiency and related myeloid neoplasms. *Semin. Hematol.* 2017;54(2):81–86.
42. Johnson KD, Kong G, Gao X, et al. Cis-regulatory mechanisms governing stem and progenitor cell transitions. *Sci. Adv.* 2015;1(8):.
43. Zhu Y, Sun L, Chen Z, et al. Predicting enhancer transcription and activity from chromatin modifications. *Nucleic Acids Res.* 2013;41(22):10032–10043.
44. Margueron R, Reinberg D. The Polycomb complex PRC2 and its mark in life. *Nature*. 2011;469(7330):343–349.
45. Menendez-Gonzalez JB, Vukovic M, Abdelfattah A, et al. Gata2 as a Crucial Regulator of Stem Cells in Adult Hematopoiesis and Acute Myeloid Leukemia. *Stem Cell Reports*. 2019;13(2):291–306.

46. Zhang SJ, Ma LY, Huang QH, et al. Gain-of-function mutation of GATA-2 in acute myeloid transformation of chronic myeloid leukemia. *Proc. Natl. Acad. Sci. U. S. A.* 2008;105(6):2076–2081.
47. Kazenwadel J, Secker GA, Liu YJ, et al. Loss-of-function germline GATA2 mutations in patients with MDS/AML or MonoMAC syndrome and primary lymphedema reveal a key role for GATA2 in the lymphatic vasculature. *Blood*. 2012;119(5):1283–1291.
48. Spinner MA, Sanchez LA, Hsu AP, et al. GATA2 deficiency: A protean disorder of hematopoiesis, lymphatics, and immunity. *Blood*. 2014;123(6):809–821.
49. Hsu AP, Johnson KD, Falcone EL, et al. GATA2 haploinsufficiency caused by mutations in a conserved intronic element leads to MonoMAC syndrome. *Blood*. 2013;121(19):3830–3837.
50. Luesink M, Hollink IHIM, van der Velden VHJ, et al. High GATA2 expression is a poor prognostic marker in pediatric acute myeloid leukemia. *Blood*. 2012;120(10):2064–2075.
51. Vicente C, Vazquez I, Conchillo A, et al. Overexpression of GATA2 predicts an adverse prognosis for patients with acute myeloid leukemia and it is associated with distinct molecular abnormalities. *Leukemia*. 2012;26(3):550–554.
52. Celton M, Forest A, Gosse G, et al. Epigenetic regulation of GATA2 and its impact on normal karyotype acute myeloid leukemia. *Leukemia*. 2014;28(8):1617–1626.
53. Wouters BJ, Löwenberg B, Erpelinck-Verschueren CAJ, et al. Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome. *Blood*. 2009;113(13):3088–3091.
54. Dufour A, Schneider F, Metzeler KH, et al. Acute myeloid leukemia with biallelic CEBPA gene mutations and normal karyotype represents a distinct genetic entity associated with a favorable clinical outcome. *J. Clin. Oncol.* 2010;28(4):570–577.
55. Fasan A, Haferlach C, Alpermann T, et al. The role of different genetic subtypes of CEBPA mutated AML. *Leukemia*. 2014;28(4):794–803.
56. Leubolt G, Redondo Monte E, Greif PA. *GATA2* mutations in myeloid malignancies: Two zinc fingers in many pies. *IUBMB Life*. 2020;72(1):151–158.
57. Batcha AMN, Bamopoulos SA, Kerbs P, et al. Allelic Imbalance of Recurrently Mutated Genes in Acute Myeloid Leukaemia. *Sci. Rep.* 2019;9(1):11796.
58. Al Seraihi AF, Rio-Machin A, Tawana K, et al. GATA2 monoallelic expression underlies reduced penetrance in inherited GATA2-mutated MDS/AML. *Leukemia*. 2018;32(11):2502–2507.
59. Kozyra EJ, Pastor VB, Lefkopoulos S, et al. Synonymous GATA2 mutations result in selective loss of mutated RNA and are common in patients with GATA2 deficiency. *Leukemia*. 2020;34(10):2673–2687.
60. Ottema S, Mulet-Lazaro R, Beverloo HB, et al. Atypical 3q26/MECOM rearrangements genocopy inv(3)/t(3;3) in acute myeloid leukemia. *Blood*. 2020;136(2):224–234.
61. Suzuki M, Katayama S, Yamamoto M. Two effects of GATA2 enhancer repositioning by 3q chromosomal rearrangements. *IUBMB Life*. 2020;72(1):159–169.
62. Saida S, Zhen T, Kim E, et al. Gata2 deficiency delays leukemogenesis while contributing to aggressive leukemia phenotype in Cbfb-MYH11 knockin mice. *Leukemia*. 2020;34(3):759–770.
63. Di Genua C, Valletta S, Buono M, et al. C/EBPα and GATA-2 Mutations Induce Bilineage Acute Erythroid Leukemia through Transformation of a Neomorphic Neutrophil-Erythroid Progenitor. *Cancer Cell*. 2020;37(5):690-704.e8.
64. Horsthemke B. Epimutations in human disease. *Curr. Top. Microbiol. Immunol.* 2006;310:45–59.
65. Plass C, Pfister SM, Lindroth AM, et al. Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. *Nat. Rev. Genet.* 2013;14(11):765–780.

SUPPLEMENTARY INFORMATION

SUPPLEMENTARY METHODS

Patient material

Samples of *de novo* AML patients were collected from the biobanks of the Erasmus MC Hematology department (Rotterdam, The Netherlands) and the University Hospital Regensburg Internal Medicine department (Regensburg, Germany). Mononuclear cells were isolated from bone marrow or peripheral blood as described previously[1]. Nine CD34+ bone marrow cells from healthy allogenic donors were obtained from the Erasmus MC bone marrow transfer unit, and three CD34+ cord blood cells from healthy infants were obtained from the University Hospital Regensburg. All patients provided written informed consent in accordance with the Declaration of Helsinki. Patient blasts were stored at -80°C in RLT+ buffer (Qiagen) and RNA and DNA was isolated using the AllPrep DNA/RNA mini kit (Qiagen, #80204) or stored in RNABee (Tel-Test, Inc.) and isolated by standard diagnostic procedures. RNA was converted into cDNA using the SuperScript II Reverse Transcriptase (Thermo Fischer Scientific) according to standard diagnostic procedures.

RNA sequencing

Sample libraries were prepped using 500 ng of input RNA according to the KAPA RNA HyperPrep Kit with RiboErase (HMR) (Roche) using Unique Dual Index adapters (Integrated DNA Technologies, Inc.). Amplified sample libraries were paired-end sequenced (2x100 bp) on the Novaseq 6000 platform (Illumina) and aligned against the human genome (hg19) using STAR v2.5.4b[2].

Whole exome sequencing

The Genomic DNA Clean & Concentrator kit (ZYMO Research) was used to remove EDTA from the DNA samples. Sample libraries were prepared using 100 ng of input according to the KAPA HyperPlus Kit (Roche) using Unique Dual Index adapters (Integrated DNA Technologies, Inc.). Exomes were captured using the SeqCap EZ MedExome (Roche Nimblegen) according to SeqCap EZ HyperCap Library v1.0 Guide (Roche) with the xGen Universal blockers – TS Mix (Integrated DNA Technologies, Inc.). The amplified captured sample libraries were paired-end sequenced (2x100 bp) on the Novaseq 6000 platform (Illumina) and aligned to the hg19 reference genome using the Burrows-Wheeler Aligner (BWA)[3], v0.7.15-r1140.

Allele-specific expression

In most ASE studies, DNA-seq and RNA-seq reads carrying reference and alternative alleles are counted as heterozygous single nucleotide variants (SNVs). Analysis of small insertions or deletions (indels) is technically very challenging and is rarely attempted, since alignment

errors over loci are pervasive[4]. In addition, a critical step is the correct identification of true heterozygous variants, which is hampered by intrinsic shortcomings of short read sequencing: sequencing errors, mapping to repetitive regions and other technical biases. Lastly, analyses must account for non-pathogenic processes leading to ASE, such as X-chromosome inactivation, imprinting and random monoallelic expression, which occur naturally in all cells and are necessary for their normal function[5].

To discriminate expression from different alleles, single nucleotide variants (SNVs) were first detected at the DNA level. This step was performed with a custom script that integrated variants called by multiple software tools, including HaplotypeCaller and MuTecT2 from GATK[6], VarScan2[7] and bcftools[8]. The combined list of SNVs was subjected to stringent filtering to remove low-quality positions, considering the following criteria: a) strand bias, b) sequencing depth, c) alignment and base calling score, d) mappability. A highly optimized in-house tool (*annotateBamStatistics*) was then used to compute DNA and RNA allele-specific read counts at every SNV position from their respective alignment (BAM) files. Positions with fewer than 9 WES reads or 5 RNA-seq reads were excluded. For every gene, counts from all SNVs were summed to create a 2x2 contingency table (variables MAJOR/MINOR and DNA/RNA) and a $\chi 2$ test of independence was conducted. Finally, skewed expression was determined for genes with False Discovery Rate (FDR) < 0.05 and RNA minor allele frequency < 0.35. To prevent false positives derived from excessive variants at low depth, genes with <= 10 RNA-seq reads/SNP were removed. Furthermore, genes with variants per kb of exonic length (VPKE) above > 1.5 and in the 95th percentile were discarded. The results were validated by visual examination of the DNA-seq and RNA-seq BAM files in IGV[9].

After the initial exploratory screen, a targeted, manually curated analysis was conducted on *GATA2* to identify cases missed by the automated pipeline. Since the coverage of WES is low or null in UTR regions, where SNVs are often located, detection of ASE was based only on RNA-seq for samples without SNVs in any other part of the gene. For example, cases 2240 and 4336 had very low DNA coverage in the 3'-UTR of the gene, resulting in the exclusion of SNVs in that area. Moreover, ASE was determined only with RNA minor allele frequency < 0.35 for positions with more than 20 reads, without statistical testing.

Differential expression analysis

Salmon[10] was used to quantify expression of individual transcripts, which were subsequently aggregated to estimate gene-level abundances with tximport[11]. Human gene annotation derived from GENCODE[12] (v30) was downloaded as a GTF file. Both gene- and transcript-level abundances were normalized to counts per million (CPM) for visualization in the figures of this paper. Differential gene expression analysis of count estimates from Salmon was performed with DESEq2[13] (v1.24.0). For the comparison of groups with and without *GATA2* allele-specific expression, we removed samples with <1 TPM from the "no ASE" group in order to prevent them from skewing the average expression levels.

Statistical association between mutations and genes with ASE

Since we did not have control material for the studied AML patients, we selected mutations likely to be somatic among the variants identified by WES based on functional annotation by Annovar[14]. Thus, we first considered mutations complying with the following criteria: a) located in exons or in splicing acceptor regions, b) non-synonymous SNV or indels, c) with a VAF of at least 5%, d) previously reported in AML or relevant for myeloid development. Single nucleotide polymorphisms (SNPs) with a population frequency higher than 0.01 were excluded unless they were reported in the COSMIC database v88, in at least 5 hematological cancers[15]. Variants present in a healthy donor (though not a paired matched control) were also removed to further eliminate common variants and technical artifacts. Finally, probable oncogenic variants were selected as those that fulfilled one or more of the following conditions: i) in COSMIC database; ii) frameshift, stopgain or startloss; iii) majority of damaging functional predictions by tools such as PolyPhen, SIFT, LRT and others. The list was further validated by manual inspection.

We compiled the mutations identified in all patients into a matrix where samples are rows and genes are columns (Supplementary Table S3). Subsequently, we calculated the statistical association between every possible pair of mutated genes and genes with ASE based on the co-occurrence of these two events in the patient cohort, using a Fisher's exact test. The results were presented as a heatmap, where the -log10(p-value) was multiplied by the sign of the odds ratio, in such a way that high positive values correspond to positive associations, whereas negative values correspond to negative associations.

For descriptive statistics and hypotheses tests involving clinical variables, the R package *Atable*[16] was used with customized settings and functions. Statistical tests involving numerical variables (e.g. age) were computed with a Kolmogorov-Smirnov test and the effect size was measured with Cohen's D. Categorical variables (e.g. gene mutations) were tested with a Fisher's exact test and the effect size was measured as odds ratio.

TCGA and Beat AML cohorts

We obtained authorization from the TCGA and the Beat AML consortia to download and analyze their data. Since they were only used for validation of *GATA2* ASE in *CEBPA* DM cases, the datasets were downloaded in BAM format and not realigned afterwards.

Following the same criteria used in our in-house cohort, we defined CEBPA DM as samples with either 2 CEBPA mutations or a single mutation with VAF > 50% (homozygous). In the TCGA cohort, we detected 19 samples with CEBPA mutations, 5 of which were bona fide CEBPA DM. However it is likely there are more because *CEBPA* is difficult to sequence due its high CG content, and previous studies generally report a similar number of single and double mutant *CEBPA*[17,18]. In the Beat AML cohort, we detected 15 cases complying with our definition. Of these 15 cases, data from DNA and RNA was only available to us for 6 cases, two of which had *GATA2* mutations.

Amplicon sequencing

Custom amplicon sequencing panels were designed for the 3 major isoforms for both DNA and cDNA level using DesignStudio Sequencing Assay Designer (Illumina) (See Supplementary Table S6). A nested PCR was used to generate the libraries. The first PCR was used to generate the amplicons using the Q5 High-Fidelity DNA Polymerase (New England Biolabs, NEB) with the following PCR program: 98°C 5 minutes, 98°C 30 seconds, 64°C for cDNA/69°C for DNA 30 seconds, steps 2 and 3 were repeated for 25 cycles, followed by 72°C for 7 minutes. A 0,8x AMPure XP Bead cleanup (Beckman Coulter) was done according to the manufacturer□s protocol before the adapters were ligated in the second PCR using the KAPA HiFi Hotstart Ready Mix (Roche) with the following PCR program: 95°C 5 minutes, 98°C 20 seconds, 66°C 30 seconds, 72°C 30 seconds, steps 2, 3 and 4 were repeated for 10 cycles, followed by 72°C for 1 minute. The amplified libraries were cleaned up using a 1,1x AMPure XP Bead cleanup according to the manufacturer's protocol followed by paired-end sequencing (2x250 bp) using the MiSeq platform (Illumina) and aligned to the hg19 reference genome.

ERRBS data and analysis

ERRBS data previously published by our group were retrieved from public repositories[19]. Raw aligned reads and methylated base calls for CpGs were imported, filtered and normalized with the package *methylKit*[20] (v1.13.1). *CEBPA* silenced leukemias were excluded from the analysis due to their strongly methylated profile, which sets them apart from any other AML subtype [21,22]. Promoter regions were defined as the 1000 bp region upstream of the transcriptional start site (TSS) of transcripts in the Ensembl gene annotation, which was downloaded with the biomaRt package[23]. Comparisons across groups of interest (*CEBPA* DM, AML with *GATA2* ASE and without) were performed with *methylKit* and average methylation levels were plotted along the *GATA2* gene with *Gviz*[24] (v1.28.3). AML samples with rearrangements involving the 3q21 region were excluded because in that AML subtype *GATA2* ASE is due to loss of the distal -110 kb *GATA2* super-enhancer[25].

Targeted bisulfite DNA amplicon sequencing

For the bisulfite conversion of genomic DNA, the EZ DNA methylation kit (Zymoresearch D5001) was used. Purified, converted DNA was used to generate amplicon libraries with a 2-step PCR amplification. In the step 1 PCR, the GATA2 promoter regions were amplified using primers that contain the GATA2 promoters and the Illumina sequencing primers. Amplification was performed using EpiMark Hot Start Taq DNA Polymeras (NEB), 0.2 µM each primer, and 2 µl of purified, converted genomic DNA. Cycling conditions were initial denaturation at 95°C for 30 s, 40 cycles of denaturation at 95°C for 30 s, annealing at 54°C for 60 s, and elongation at 68°C for 60 s, followed by a final elongation step at 68°C for 5 min. The amplicons were cleaned up with 0.8x AMPure XP (Agencourt/Beckman Coulter) beads

by following the manufacturer's instructions. The step 2 PCR targets the Illumina sequencing primer to add the indices. Amplification was performed using EpiMark Hot Start Taq DNA Polymeras (NEB) with 0.4 µM of each primer and 10 ng purified amplicon. Cycling conditions were initial denaturation at 95°C for 5 min, 10 cycles of denaturation at 95°C for 20 s, annealing at 66°C for 30 s, and elongation at 68°C for 60 s, followed by a final elongation step at 68°C for 5 min. The amplicons were cleaned up with 1.1x AMPure XP (Agencourt/ Beckman Coulter) beads according to the manufacturer's instructions. Sequencing was done using the Illumina MiSeq platform.

Oxford Nanopore sequencing and analysis

CRISPR/Cas9-targeted enrichment followed by amplification-free long read sequencing by Oxford Nanopore were performed as previously described[21]. Briefly, genomic DNA isolated from fresh frozen samples was dephosphorylated and crRNAs were used to target Cas9 to the *GATA2* locus. Following the introduction of double-strand breaks, Oxford Nanopore sequencing adapters were ligated and the resulting libraries were sequenced on Flongle flow cells. Two crRNA sequential guides were designed for each end of the targeted region, namely:

3' exonic end: TGGCTGGCATCGTGTTCCCC, GCCCAGGGCCCCAGGCGTTC

5' promoter end: TTTCACTCTCCCTGATCTGC, AAGCCCAAGCACTTTCCCTC

Nanopore reads were aligned with bwa using default settings. The reads in the targeted region spanned 10 kb of the *GATA2* gene, ranging from Prom-L to the third exon and including at least one heterozygous SNP to enable distinction between two alleles. The reads were thus assigned to one of two alleles based on this SNP, previously detected in the WES data: rs1573858 (C/G) for 2253, 2273 and 3327 and rs72983369 (A/G) for 2240. The alleles were labeled as either active (A) or inactive (I) depending on whether they were expressed, according to RNA-seq data.

A methylation log likelihood ratio (modified base/canonical base) was computed with Nanopolish for every CpG position in the captured region[26]. Average log likelihood ratios (LLR) of all the reads in either the A or I alleles were computed across the entire region of interest, and plotted using the R package Gviz[24]. This was done for each patient separately, effectively linking methylation levels and allele expression in every case. Furthermore, the data were aggregated in the putative promoter regions defined for the bisulfite sequencing analysis (Prom-L, Prom-S) and the alleles were compared with a Wilcoxon rank-sum test. For the methylation profile depicted in Figure S9A, LLR of individual reads were converted into discrete methylation values as follows: 100 if LLR > 1, 50 if -1 < LLR < 1, 0 if LLR < -1. When aggregated across all reads, these values result in frequences resembling data from bisulfite sequencing.

In addition to Nanopolish, other methylation callers were tested in patient 3327, including Megalodon and Guppy. They all showed high consistency in their methylation calls, as shown in Figure S9B.

ChIP-seq data and analysis

ChIP-seq data were generated for a number of selected patients to investigate changes in enhancer and promoter regions. ChIP was performed as described previously with slight modifications[27]. Briefly, cells were crosslinked with 1% formaldehyde for 10 minutes at room temperature and the reaction was quenched with glycine at a final concentration of 0.125 M. Chromatin was sheared using the Covaris S220 focused-ultrasonicator to an average size of 250–350 bp. A total of 2.5 µg of antibody against H3K27ac (Abcam, ab4729) was added to sonicated chromatin of 2×10^6 cells and incubated overnight at 4 °C. Protein A sepharose beads (GE healthcare) were added to the ChIP reactions and incubated for 2 h at 4 °C. Beads were washed and chromatin was eluted. After crosslink reversal, RNase A and proteinase K treatment, DNA was extracted with the Monarch PCR & DNA Cleanup kit (NEB). Sequencing libraries were prepared with the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB) according to the manufacturer's instructions. The quality of dsDNA libraries was analyzed using the High Sensitivity D1000 ScreenTape Kit (Agilent) and concentrations were assessed with the Qubit dsDNA HS Kit (Thermo Fisher Scientific). Libraries were single-end sequenced on a HiSeq3000 (Illumina).

ChIP-seq reads were aligned to the human reference genome build hg19 with *bowtie* and bigwig files were generated for visualization with bedtools genomecov[28] (v2.27.1) and UCSC bedGraphToBigWig[29]. Peak calling was performed with MACS2[30] (v 2.1.2) using default settings.

ATAC-seq data and analysis

ATAC-seq data were generated for a number of selected patients to investigate changes in enhancer and promoter regions. ATAC-seq was essentially carried out as described in[31]. Briefly, prior to transposition the viability of the cells was assessed and 1×10^6 cells were treated in culture medium with DNase I (Sigma) at a final concentration of 200 U ml−1 for 30 minutes at 37 °C. After Dnase I treatment, cells were washed twice with ice-cold PBS, and cell viability and the corresponding cell count were assessed. 5 × 104 cells were aliquoted into a new tube and spun down at 500 × g for 5 minutes at 4 °C, before the supernatant was discarded completely. The cell pellet was resuspended in 50 µl of ATAC-RSB buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl2) containing 0.1% NP-40, 0.1% Tween-20, and 1% Digitonin (Promega), and was incubated on ice for 3 minutes to lyse the cells. Lysis was washed out with 1 ml of ATAC-RSB buffer containing 0.1% Tween-20. Nuclei were pelleted at 500 × g for 10 minutes at 4 °C. The supernatant was discarded carefully and the cell pellet was resuspended in 50 µl of transposition mixture (25 µl 2× tagment DNA buffer, 2.5 µl transposase (100 nM final; Illumina), 16.5 µl PBS, 0.5 µl 1% digitonin, 0.5 µl 10% Tween-20, 5 µl H2O) by pipetting up and down six times. The reaction was incubated at 37 °C for 30 minutes with mixing before the DNA was purified using the Monarch PCR & DNA Cleanup Kit (NEB) according to the manufacturer's instructions. Purified DNA was eluted in 20 µl elution buffer (EB) and 10 µl purified sample was objected to a ten-cycle PCR amplification

using Nextera i7- and i5-index primers (Illumina). Purification and size selection of the amplified DNA were carried out with Agencourt AMPure XP beads. For purification the ratio of sample to beads was set to 1:1.8, whereas for size selection the ratio was set to 1:0.55. Purified samples were eluted in 15 µl of EB. Quality and concentration of the generated ATAC libraries were analyzed using the High Sensitivity D1000 ScreenTape Kit (Agilent) and libraries were sequenced paired-end on a NovaSeq (Illumina).

ATAC-seq reads were aligned to the human reference genome build hg19 with *bowtie2*[32] (v2.3.4.1), which is recommended for longer reads, and mitochondrial and duplicate reads were excluded. Bigwig files were generated as described above. Peak calling was also performed with MACS2[30] (v 2.1.2), but with the following settings*: --nomodel --shift 100 --extsize 200*.

Identification and analysis of enhancer regions

Enhancer regions were defined for quantification of eRNA from RNA-seq, as well as H3K27ac, H3K27me3 and ATAC-seq reads. Therefore, they were chosen based on two complementary criteria: a) expression of eRNA, and b) active enhancers in AML. To that end, we downloaded putative enhancers detected by CAGE-seq by the FANTOM consortium[33] (*human_permissive_enhancers_phase_1_and_2.bed*) and intersected them with merged H3K27ac ChIP-seq peaks from 30 AML patients. The resulting BED file was then converted into GTF with the UCSC tools *bedToGenePred* and *genePredToGtf*[29]. Read counts in enhancer regions were computed with featureCounts[34] (v1.5.0-p3) and differential analysis was conducted with DESEq2[13] (v1.24.0). The results of this analysis were plotted in the *GATA2* region with *Gviz*[24] (v1.28.3).

SUPPLEMENTARY FIGURES

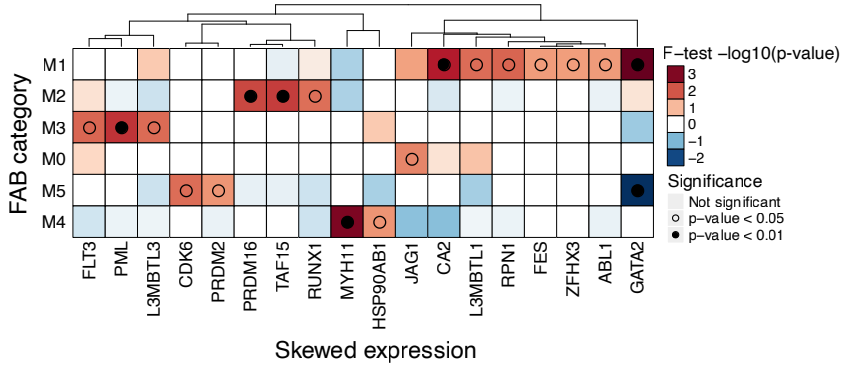


Figure S1. Association between genes with ASE and FAB classification. Statistical association was computed with a 2-sided Fisher's exact test and represented as -log10(p-value) for odds ratio > 1 or log10(p-value) for odds ratio < 1. Positive values, indicating positive association, were depicted in red, whereas negative values were depicted in blue. For a clearer visualization, the limits of the scale were set at -4 and +4. Associations that achieve significance were highlighted with an empty (p-value < 0.05) or a full (p-value < 0.01) dot.

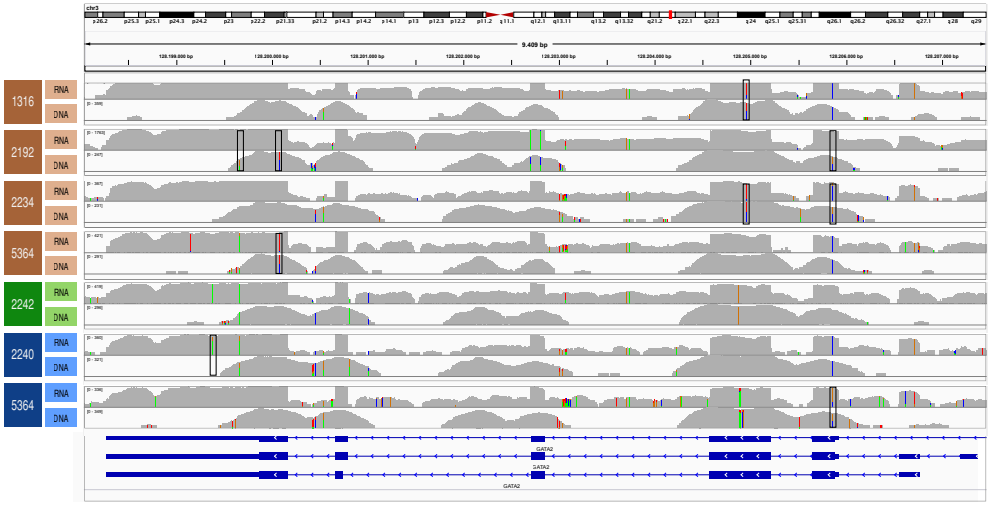


Figure S2. IGV visualizationvv nucleotide variants (SNVs) have a variant allele frequency (VAF) close to 50% in the DNA track, but one allele is preferentially expressed in the RNA. Rectangles highlight SNVs that indicate presence of ASE. In brown, *CEBPA* DM with ASE detected by the automated pipeline. In green, patient 2242, where the absence of SNV in *GATA2* makes it impossible to determine ASE. In blue, patients where *GATA2* ASE was not detected by the automated pipeline due to low coverage in exonic regions.

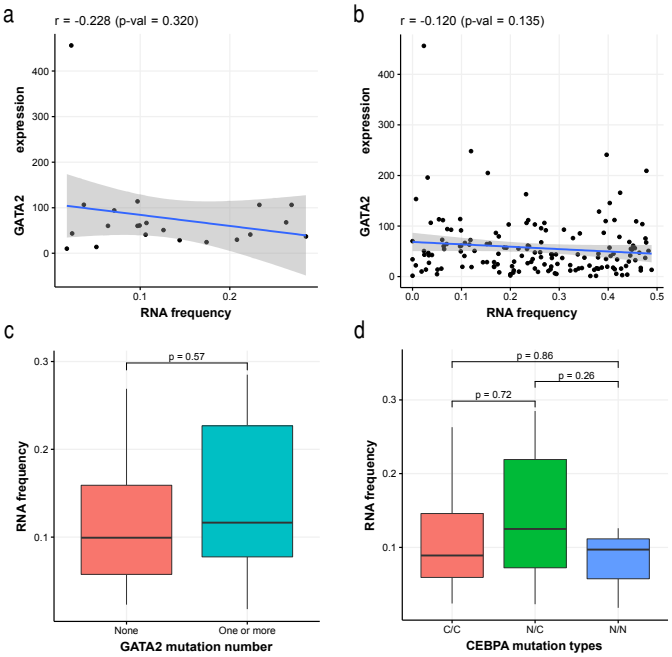


Figure S3. Associations between *GATA2* ASE and other parameters. (A) Correlation between *GATA2* ASE (measured as frequency of reads coming from the minor allele in the RNA) and *GATA2* expression in the *CEBPA* DM group. (B) Correlation between *GATA2* ASE and *GATA2* expression in the whole cohort (only patients with SNVs, n=170). (C) Association between the number of *GATA2* mutations and *GATA2* ASE (D) Association between type of *CEBPA* mutations and *GATA2* ASE

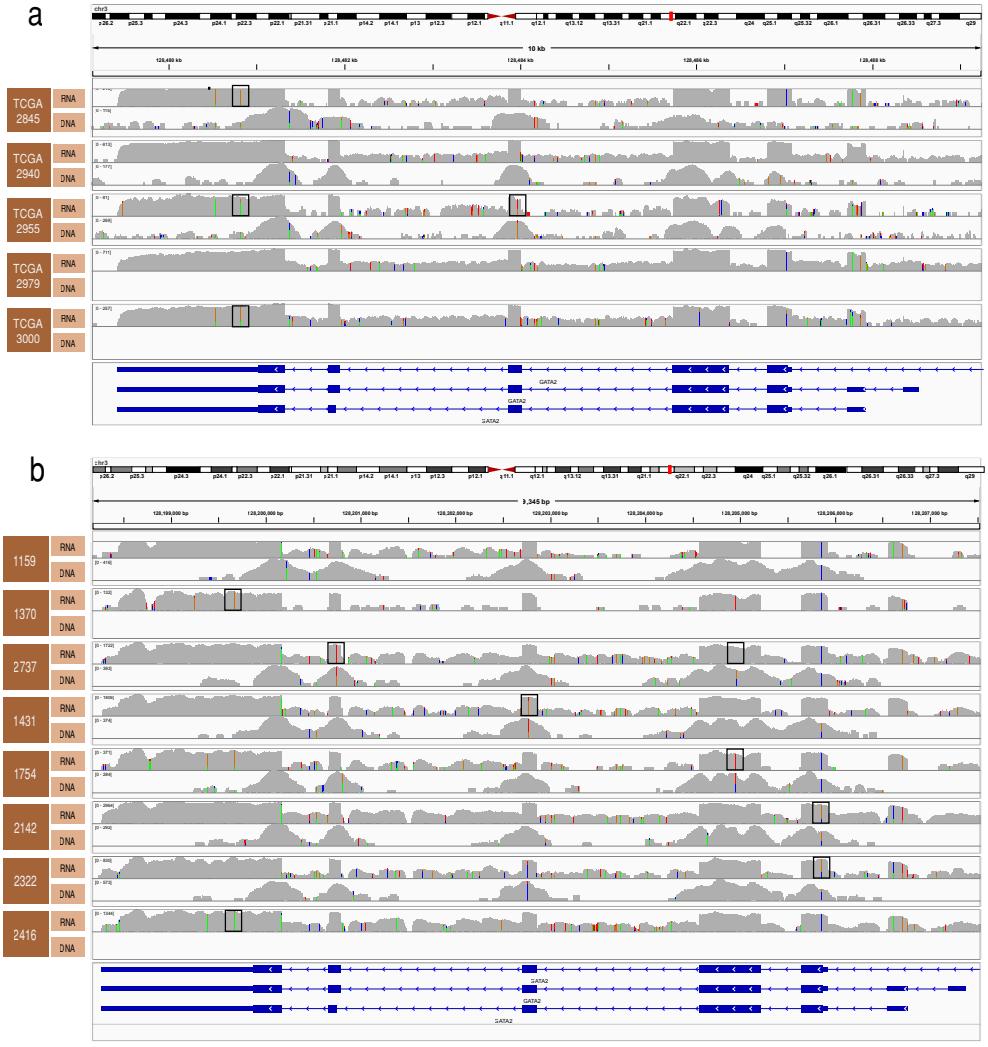


Figure S4. Confirmation of *GATA2* ASE in *CEBPA* DM patients from other cohorts. Rectangles highlight SNVs that indicate presence of ASE. (A) IGV visualization of 5 cases identified in the TCGA-LAML cohort. (B) IGV visualization of 8 cases identified in the Beat AML cohort.

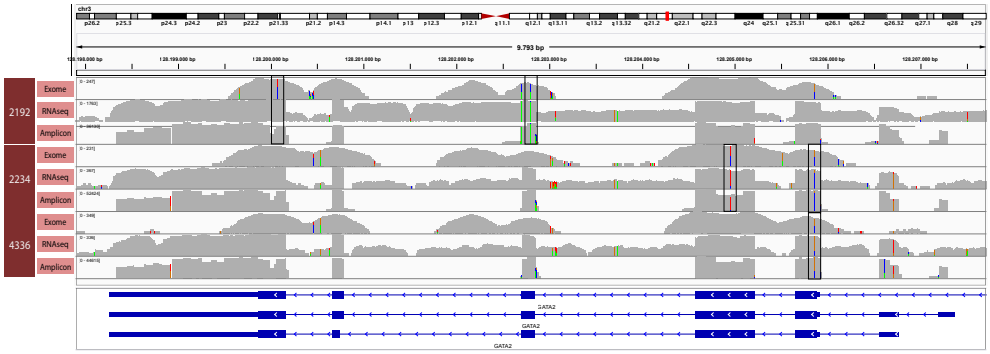


Figure S5. Comparison of RNA-seq and amplicon-seq results. IGV visualization of 3 *CEBPA* DM AML samples on which exome-seq, RNA-seq and amplicon-seq of the *GATA2* locus was done. The variant allele frequency (VAF) of SNVs indicating *GATA2* ASE is very similar for both amplicon-seq and RNA-seq.

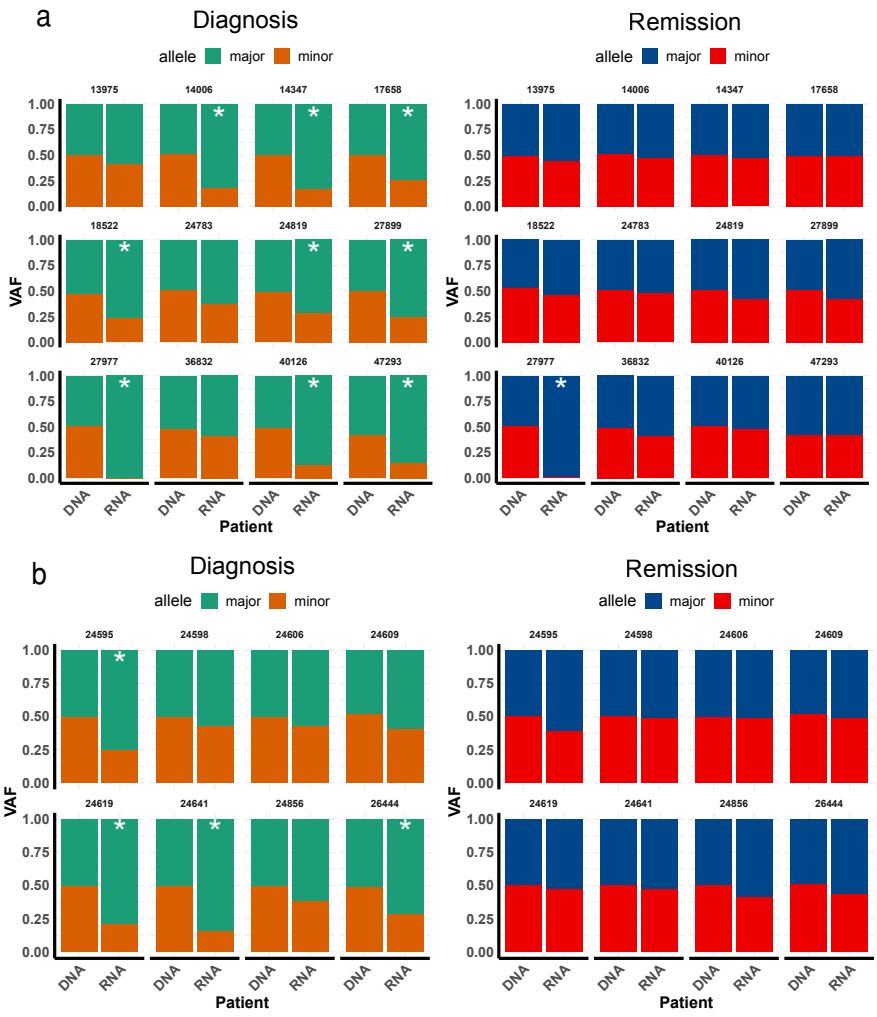


Figure S6. *GATA2* ASE detected at diagnosis is restored at remission. This figure contains the same data as Figures 4B and 4C. (A) Bar plots showing *GATA2* ASE in *CEBPA* DM AML patients at diagnosis and remission, observed by the discrepancy between VAF at the DNA level and VAF at the RNA level. (B) Bar plots showing *GATA2* ASE in *NPM1* mutated AML patients at diagnosis and remission. An asterisk (*) indicates significance at FDR < 0.05 in a chi-square test.

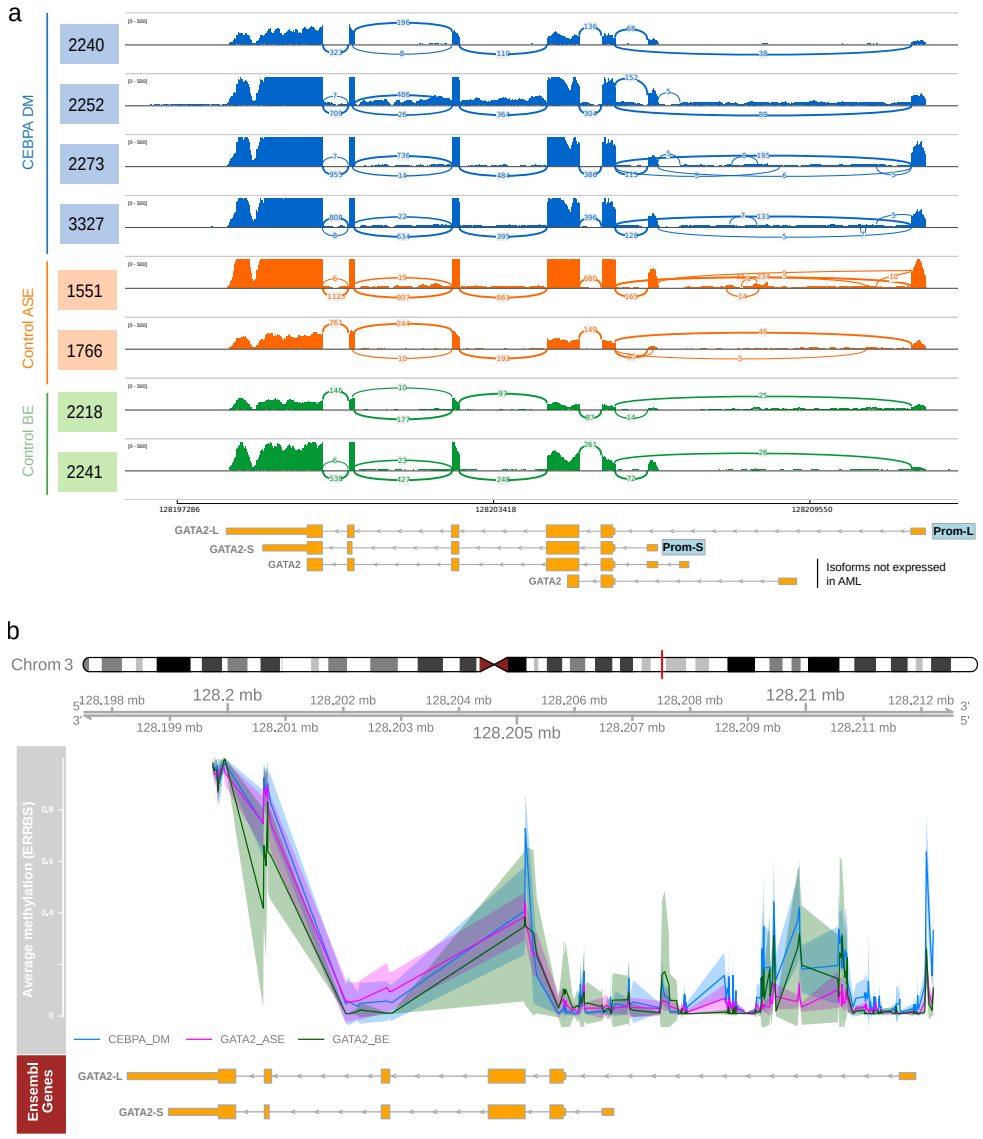


Figure S7. Methylation of *GATA2* promoter regions by bisulfite ERRBS. (A) Sashimi plot showing the expression of various *GATA2* isoforms in *CEBPA* DM AML, other AMLs with *GATA2* ASE and AMLs without *GATA2* ASE. Only two isoforms are expressed at observable levels in AML, hereafter referred to as long (L) and short (S). (B) Average methylation frequency assessed by ERRBS along the *GATA2* gene in AMLs with *CEBPA* DM (CEBPA_DM, n=10), other AMLs with *GATA2* ASE (Control_ASE, n=20) and AMLs with biallelic *GATA2* expression. The data used in this figure are the same as in Figure 5.

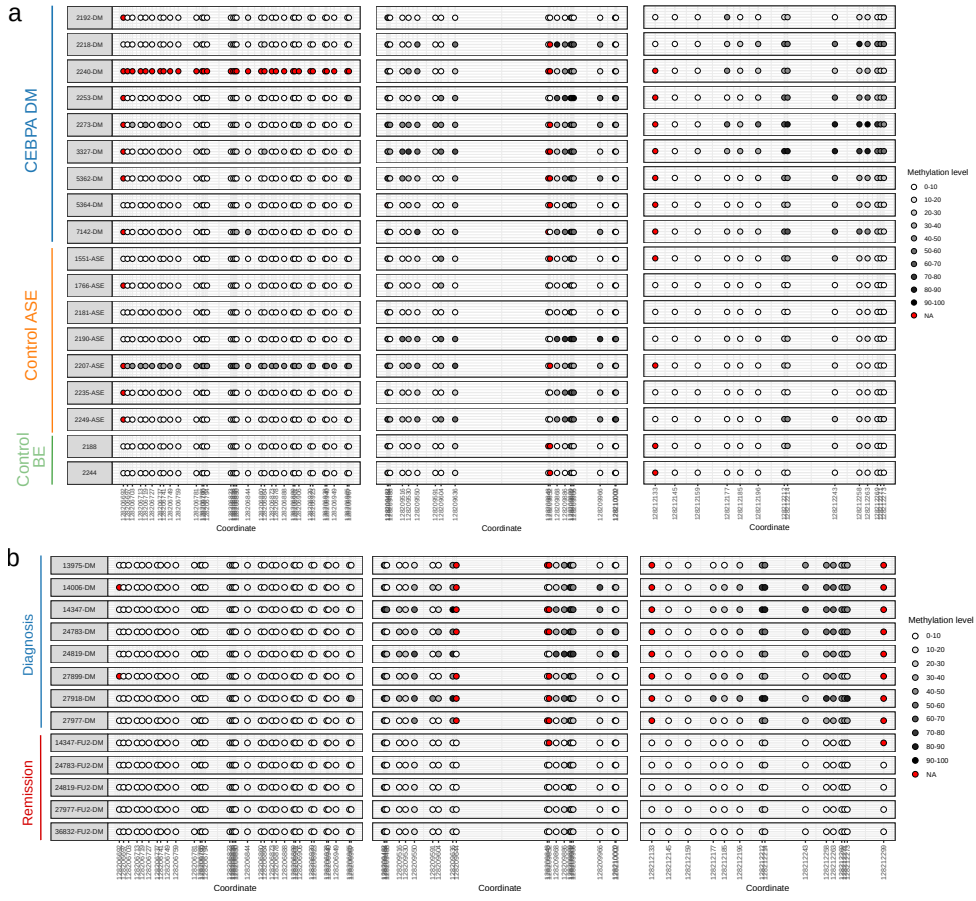


Figure S8. Methylation of *GATA2* promoter regions by bisulfite amplicon-seq. This figure contains the same data as Figure 5. (A) Lollipop plot displaying methylation levels of individual CpGs measured in *CEBPA* DM AML (CEBPA_DM, n=9), other AMLs with *GATA2* ASE (Control_ASE, n=7) and other AMLs with biallelic *GATA2* expression (Control_BE, n=2). (B) Lollipop plot displaying methylation levels of individual CpGs measured in *CEBPA* DM at diagnosis (n=8) or remission (n=5).

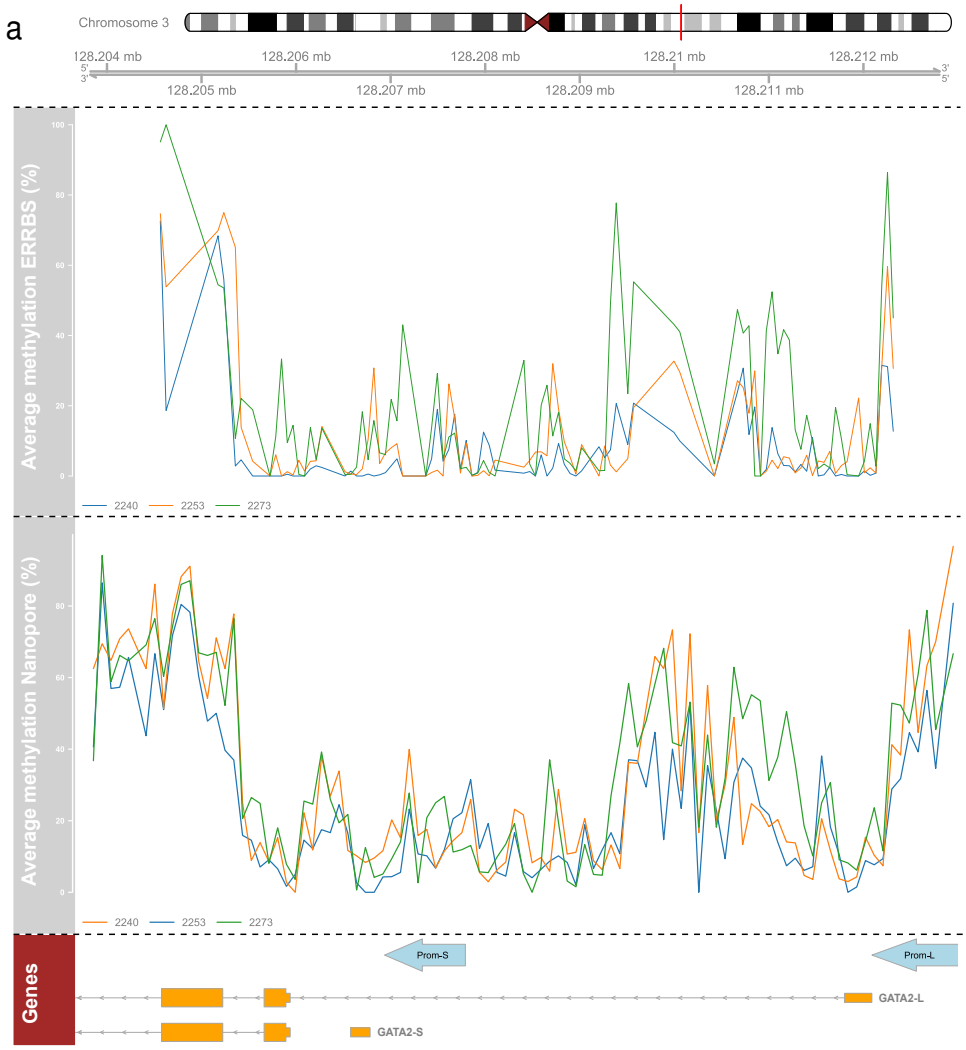
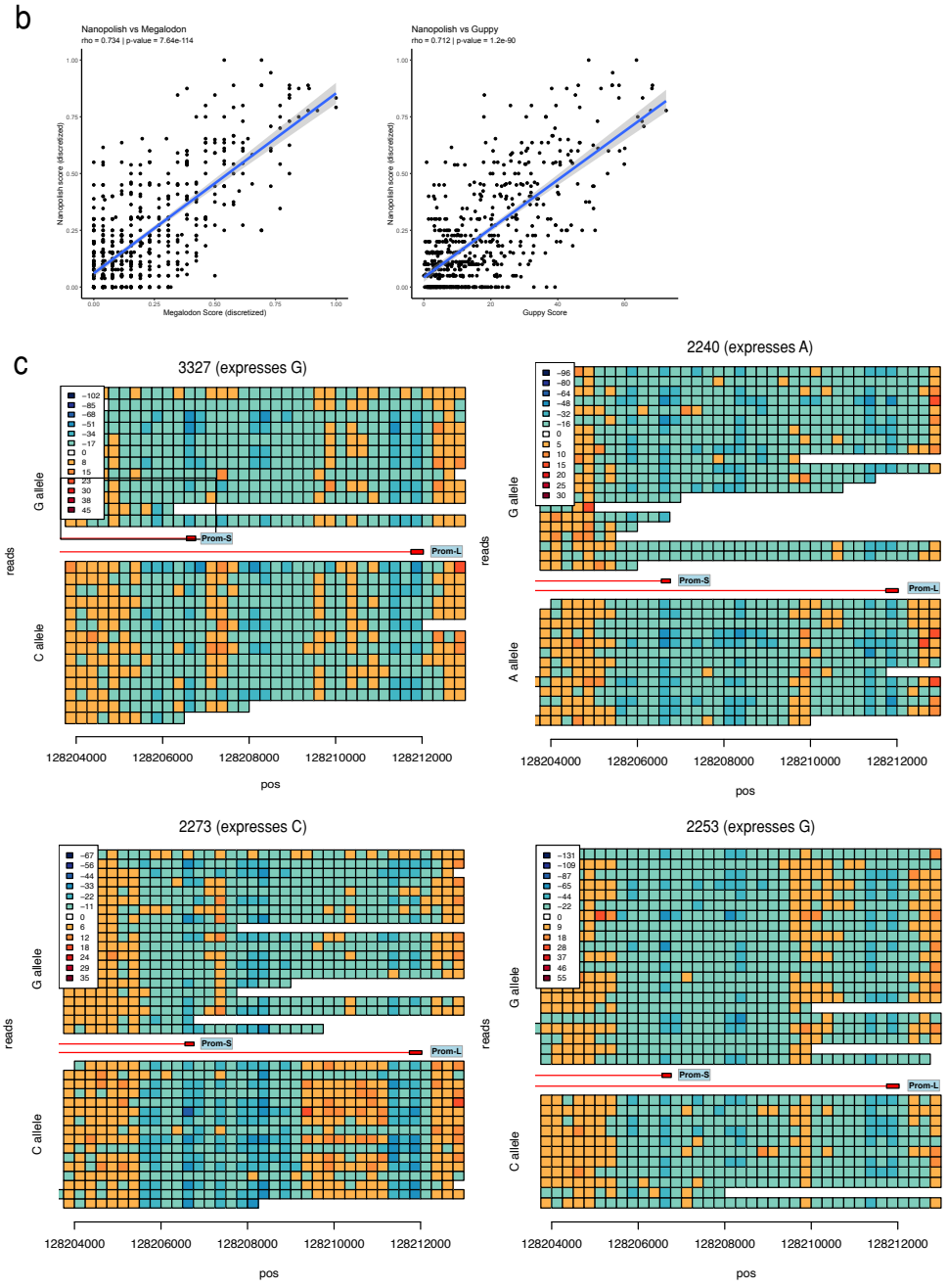


Figure S9. Allele-specific methylation in GATA2 assessed by Nanopore sequencing. (A) Comparison of ERRBS and Nanopore data in 3 *CEBPA* DM patients sequenced with both technologies. In ERRBS, the percentage corresponds to average methylation frequency across all reads, measured as C with respect to C>T conversions. In the Nanopore panel, log likelihood ratios (LLR) computed by Nanopolish were converted into discreet methylation values as follows: 100 if LLR > 1, 50 if -1 < LLR < 1, 0 if LLR < 0. The average methlation frequency was then calculated across all reads. (B) Comparison of different methylation callers in patient 3327: Nanopolish vs Megalodon in the first scatterplot and Nanopolish vs Guppy in the second (C) Heatmap displaying allele-specific methylation in 4 *CEBPA* DM, where each Nanopore read is represented as a horizontal row of 250bp bins. Bins are colored based on their LLR as calculated by Nanopolish (orange-red means methylated, green-blue means non-methylated). Putative *GATA2* promoters Prom-L and Prom-S are indicated next to their corresponding isoforms. Methylation levels of Prom-S are increased at the lowly expressed allele in patients 2253, 2273 and 3327. The data used in this figure are the same as in Figure 6.


b
Nanopolish vs Megalodon
rho = 0.734 | p-value = 7.64e-114
Nanopolish score (discretized)
Megalodon Score (discretized)
Nanopolish vs Guppy
rho = 0.712 | p-value = 1.2e-90
Nanopolish score (discretized)
Guppy Score
c
3327 (expresses G)
2240 (expresses A)
2273 (expresses C)
2253 (expresses G)
G allele
C allele
A allele
reads
Prom-S
Prom-L
128204000
128206000
128208000
128210000
128212000
pos

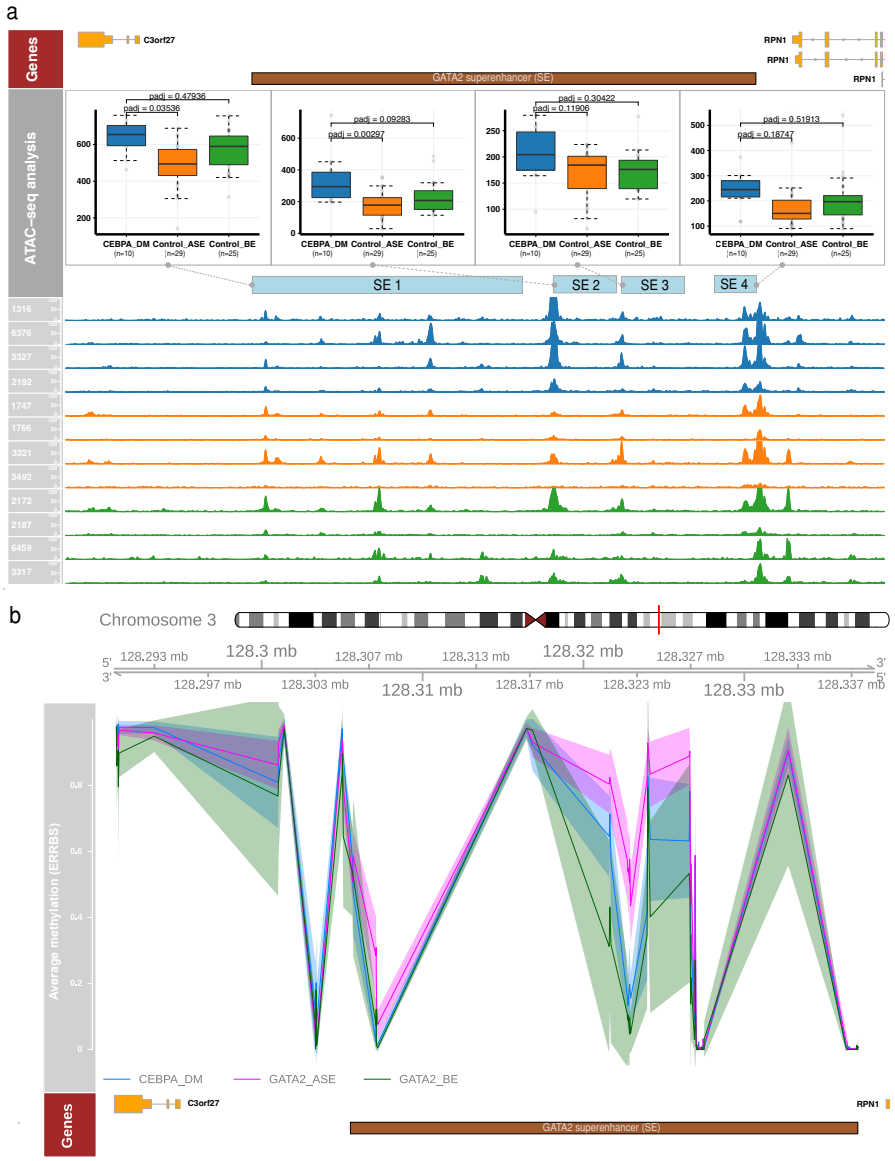
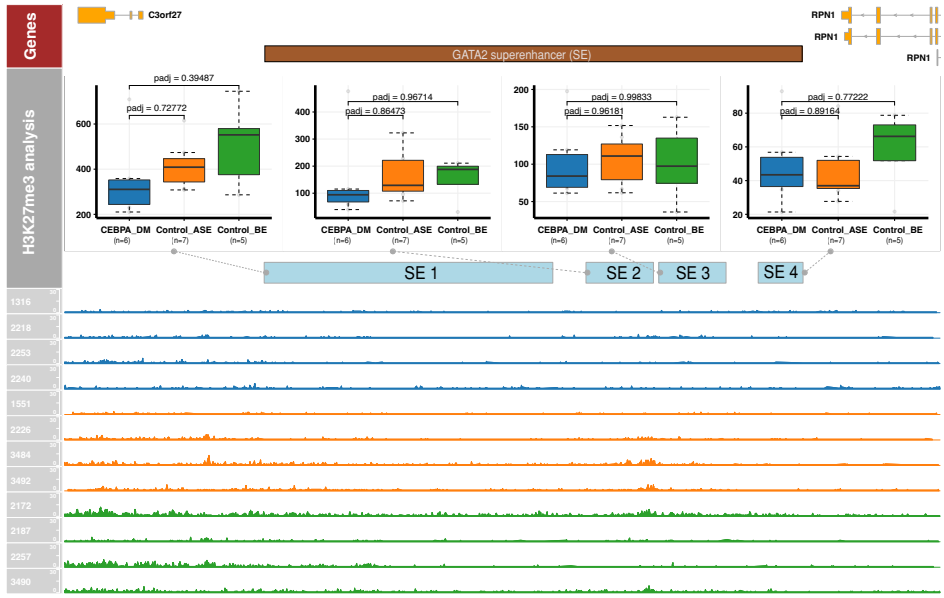
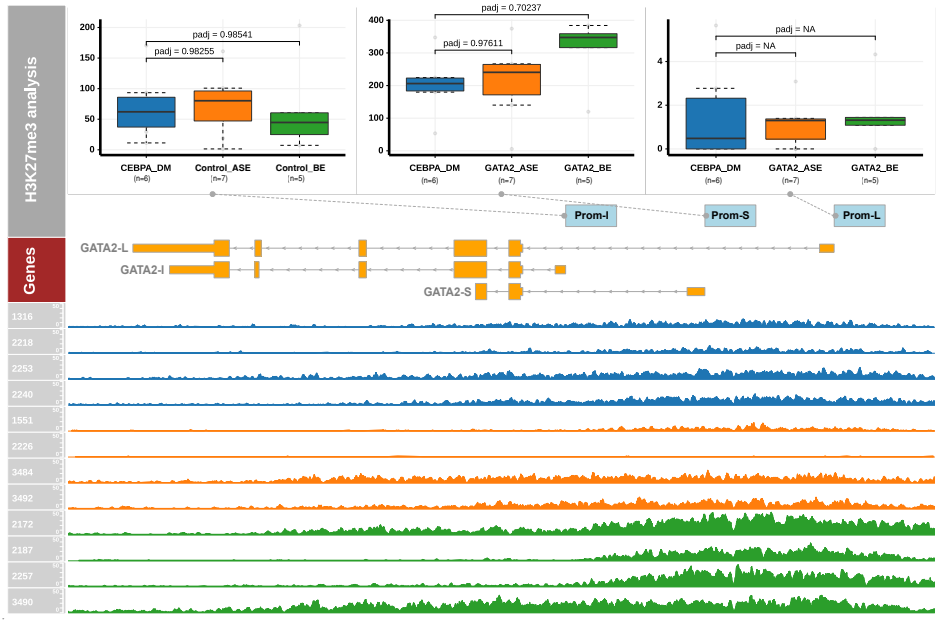


Figure S10. Analysis of *GATA2* regulatory regions. The following groups were compared: AMLs with *CEBPA* DM (CEBPA_DM), other AMLs with *GATA2* ASE (Control_ASE) and AMLs with biallelic *GATA2* expression (Control_BE). (A) Differential analysis of chromatin opening, as measured by ATAC-seq in CEBPA_DM (n=10), Control_ASE (n=29), Control_BE (n=25). (B) Methylation along the *GATA2* -110 kb super-enhancer, as measured by ERRBS. The Y-axis indicates methylation fraction, where 0 is the minimum and 1 is the maximum. Every point connected by the line is an individual CpG position -- there are few data points in this region due to the nature of the technique. (C) Analysis of H3K27me3 binding levels in the *GATA2* -110kb super-enhancer, comparing CEBPA_DM (n=6), Control_ASE (n=7) and Control_BE (n=5) (D) Analysis of H3K27me3 binding levels in the promoters of the 3 main *GATA2* isoforms (same groups as C).

c


Genes
C3orf27
RPN1
RPN1
RPN1
GATA2 superenhancer (SE)
H3K27me3 analysis
padj = 0.39487
padj = 0.72772
padj = 0.96714
padj = 0.86473
padj = 0.99833
padj = 0.96181
padj = 0.77222
padj = 0.89164
CEBPA_DM (n=6)
Control_ASE (n=7)
Control_BE (n=5)
SE 1
SE 2
SE 3
SE 4

d


H3K27me3 analysis
padj = 0.98541
padj = 0.98255
padj = 0.70237
padj = 0.97611
padj = NA
padj = NA
CEBPA_DM (n=6)
Control_ASE (n=7)
Control_BE (n=5)
GATA2_ASE (n=7)
GATA2_BE (n=5)
Prom-I
Prom-S
Prom-L
Genes
GATA2-L
GATA2-I
GATA2-S

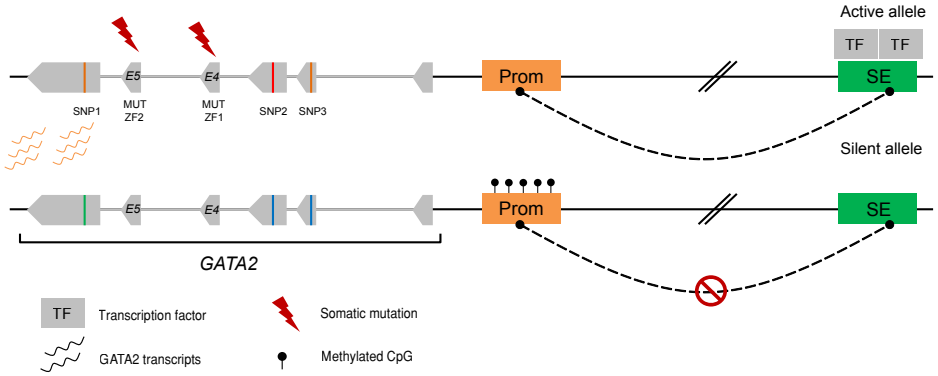


Figure S11. Proposed mechanisms for *GATA2* ASE in *CEBPA* DM AML. One allele is silenced by hypermethylation of a GATA2 promoter. The other allele is highly expressed due to the increased activation of the -110 kb super-enhancer, thus resulting in similar or even higher levels of *GATA2* compared to other AMLs.

SUPPLEMENTARY REFERENCES

1. Valk PJM, Verhaak RGW, Beijen MA, et al. Prognostically Useful Gene-Expression Profiles in Acute Myeloid Leukemia. *N. Engl. J. Med.* 2004;350(16):1617–1628.
2. Dobin A, Davis CA, Schlesinger F, et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.
3. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760.
4. Castel S, Levy-Moonshine A, Mohammadi P, Banks E, Lappalainen T. Tools and best practices for allelic expression analysis. *Genome Biolology*. 2015;16(1):195.
5. Chess A. Monoallelic Gene Expression in Mammals. *Annu. Rev. Genet.* 2016;50(1):317–327.
6. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297–1303.
7. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012;22(3):568–576.
8. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987–2993.
9. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat. Biotechnol.* 2011;29(1):24–26.
10. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*. 2017;14(4):417–419.
11. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*. 2016;4:1521.
12. Frankish A, Diekhans M, Ferreira AM, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 2019;47(D1):D766–D773.
13. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550.
14. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38(16):e164.
15. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* 2019;47(D1):D941–D947.
16. Ströbel A. atable: Create Tables for Clinical Trial Reports.
17. Konstandin NP, Pastore F, Herold T, et al. Genetic heterogeneity of cytogenetically normal AML with mutations of CEBPA. *Blood Adv.* 2018;2(20):2724–2731.
18. Fasan A, Haferlach C, Alpermann T, et al. The role of different genetic subtypes of CEBPA mutated AML. *Leukemia*. 2014;28(4):794–803.
19. Glass JL, Hassane D, Wouters BJ, et al. Epigenetic identity in AML depends on disruption of nonpromoter regulatory elements and is affected by antagonistic effects of mutations in epigenetic modifiers. *Cancer Discov.* 2017;7(8):868–883.
20. Akalin A, Kormaksson M, Li S, et al. MethylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* 2012;13(10):1–9.
21. Figueroa ME, Wouters BJ, Skrabanek L, et al. Genome-wide epigenetic analysis delineates a biologically distinct immature acute leukemia with myeloid/T-lymphoid features. *Blood*. 2009;113(12):2795–804.

22. Gebhard C, Glatz D, Schwarzfischer L, et al. Profiling of aberrant DNA methylation in acute myeloid leukemia reveals subclasses of CG-rich regions with epigenetic or genetic association. *Leukemia*. 2019;33(1):26–36.
23. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. *Nat. Protoc.* 2009;4(8):1184–1191.
24. Hahne F, Ivanek R, Lalonde E, et al. Visualizing Genomic Data Using Gviz and Bioconductor. *Source Code Biol. Med. 2016 111*. 2016;1418(43):335–351.
25. Gröschel S, Sanders MA, Hoogenboezem R, et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell*. 2014;157(2):369–381.
26. Simpson JT, Workman RE, Zuzarte PC, et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods*. 2017;14(4):407–410.
27. Pham TH, Benner C, Lichtinger M, et al. Dynamic epigenetic enhancer signatures reveal key transcription factors associated with monocytic differentiation states. *Blood*. 2012;119(24):e161–e171.
28. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–842.
29. Speir ML, Zweig AS, Rosenbloom KR, et al. The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.* 2016;44(D1):D717-25.
30. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 2008;9(9):R137.
31. Corces MR, Trevino AE, Hamilton EG, et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods*. 2017;14(10):959–962.
32. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. 2012;9(4):357–359.
33. Andersson R, Gebhard C, Miguel-Escalada I, et al. An atlas of active enhancers across human cell types and tissues. *Nature*. 2014;507(7493):455–461.
34. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923–930.

CHAPTER 7

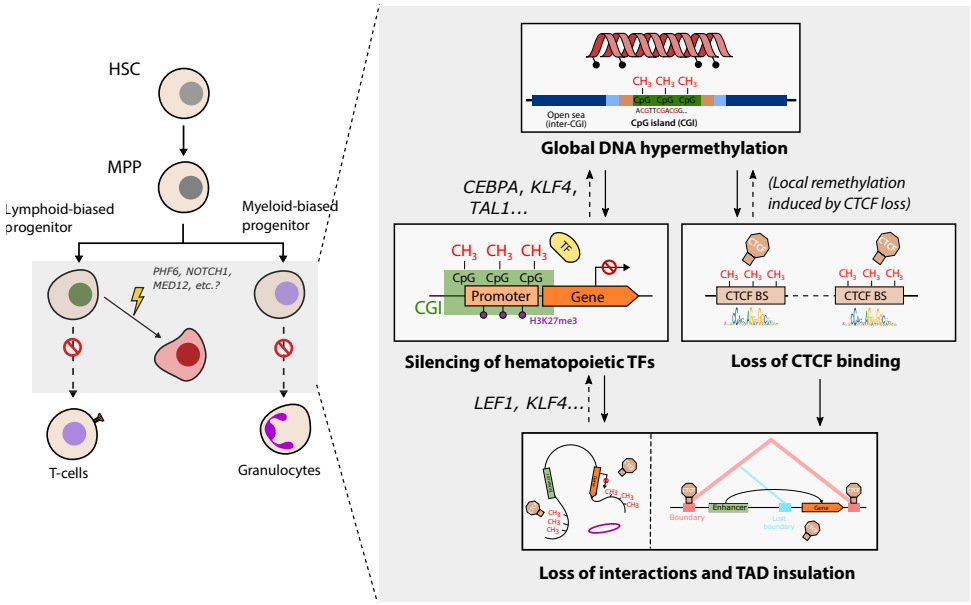
Common epigenetic signature defines mixed myeloid/lymphoid leukemias resembling ETP-ALL

Roger Mulet-Lazaro[1,2,*], Stanley van Herk[1,2], Aniko Szabo[1], Anita Rijneveld[1], Lucia Schwarzfischer[3,4], Noelia Díaz[5], Dagmar Glatz[3,4], Daniel Heudobler[3,4], Sandra Pohl[3,4], Reinhard Andreesen[3,4], Wolfgang Herr[3,4], Gerhard Ehninger[3,4], Juan M Vaquerizas[5,6,7], Christian Thiede[6], Bas Wouters[1,2], Ruud Delwel[1,2], Michael Rehli[3,4], Claudia Gebhard[3,4]

[1] Department of Hematology, Erasmus MC Cancer Institute, Rotterdam, the Netherlands
[2] Oncode Institute, Utrecht, the Netherlands
[3] Department of Internal Medicine III, University Hospital Regensburg, 93053, Regensburg, Germany
[4] Regensburg Center for Interventional Immunology (RCI), University Regensburg and University, Germany
[5] Max Planck Institute for Molecular Biomedicine, Muenster, Germany
[6] Medizinische Klinik und Poliklinik I, Universitätsklinikum Carl Gustav Carus, Dresden, Germany
[7] MRC London Institute of Medical Sciences, London, United Kingdom
[8] Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, Hammersmith Hospital 8 Campus, London, United Kingdom

Running title: Epigenetic landscape of mixed phenotype leukemias

Work in progress



ABSTRACT

Leukemias with ambiguous lineage comprise a number of loosely defined entities, often without a clear mechanistic basis. Here, we investigated a group of such leukemias with a CpG Island Methylator Phenotype (CIMP), previously identified as *CEBPA*-silenced AML. Transcriptomics and epigenomics analyses revealed a hybrid myeloid/lymphoid epigenetic landscape, whereas genetic alterations were heterogenous. This suggests that CIMP leukemias are defined by their shared epigenetic profile rather than a common genetic lesion. Gene expression enrichment suggested strong similarity with ETP-ALL and an early lymphoid progenitor cell of origin. Accordingly, integration of differential methylation and expression showed widespread silencing of myeloid transcription factors (TFs), among which *CEBPA* was key for differentiation arrest. Hypermethylation also resulted in loss of CTCF binding, accompanied by a few changes in chromatin interactions involving critical TFs like *KLF4*. In conclusion, epigenetic dysregulation, and not genetic lesions, explain the mixed phenotype of a group of CIMP leukemias resembling ETP-ALL.

STATEMENT OF SIGNIFICANCE

This study charts the epigenomic landscape underlying the mixed phenotype of a group of leukemias very similar to ETP-ALL. Moreover, the data collected here constitute a useful epigenomic reference for subsequent of acute leukemias

INTRODUCTION

Research on the pathogenesis of leukemia has traditionally emphasized the role of genetic lesions, but the importance of epigenetic dysregulation is becoming increasingly recognized. Several epigenetic modulators are recurrently mutated in acute myeloid leukemia (AML) and T-cell acute lymphocytic leukemia (T-ALL), including methylation regulators (*DNMT3A*, *TET2*, *IDH1/2*) and histone writers (*EZH2*, *SUZ12, KMT2A, KDM6A*) [1,2]. On the other hand, numerous instances of epigenetic dysregulation leading to aberrant expression of proto-oncogenes have been documented, such as the enhancer hijacking leading to *EVI1* overexpression in 3q26-rearranged AML [3,4] or the formation of a super-enhancer driving TAL1 upregulation in T-ALL [5]. However, recurrent epigenetic events may occur independently of a known genetic lesion, possibly due to selection of clones that spontaneously acquire these alterations. For example, hypermethylation of *DNMT3A* recapitulates the effects of mutations in this gene [6].

Therefore, genetic characterization of leukemia may be insufficient to identify critical pathogenic mechanisms. Accordingly, clustering of AML samples by gene expression reveals subgroups that share known genetic lesions, but also others that cannot be linked to any known abnormality [7]. One of such subgroups was later found to be defined by *CEBPA* silencing due to hypermethylation [8]. This "*CEBPA*-silenced" cluster exhibited a mixed myeloid/T-lymphoid phenotype, resistance to myeloid growth factors and possibly poor prognosis. Subsequent analyses revealed a genome-wide hypermethylation signature that distinguished this subgroup from both AML and T-cell acute lymphocytic leukemia (T-ALL), yet no mutations typically associated with methylation defects [9]. More recently, other research groups identified an AML subtype with similar characteristics and methylation localized to CpG islands (CGIs), labeling it as "CpG Island Methylator Phenotype" (CIMP) [10,11]. We hypothesize that "CIMP" and "CEBPA-silenced" leukemias are the same entity.

Hypermethylation of CGIs is a frequent event in cancer that often results in silencing of tumor suppressor genes [12]. Indeed, specific events of hypermethylation have been reported in AML, as described before. Although DNA methylation is traditionally associated with transcriptional repression, its cellular functions are in fact much more complex [13,14]. Transcriptional repression in the presence of methylation is thought to be a result of a) impaired TF binding, and b) recruitment of chromatin remodelers via methyl binding domain (MBD) proteins [13]. However, a plethora of TFs have shown the ability to bind methylated sequences [15], whereas other transcriptional regulators may be repelled by DNA methylation. A notable example of the latter is CTCF, which plays critical roles as insulator, transcriptional repressor or activator and architectural protein [16]. Thus, aberrant methylation can disrupt CTCF-dependent boundaries of topologically associating domains (TADs), resulting in dysregulated expression of neighboring genes [17,18].

Leukemias with ambiguous lineage pose substantial challenges for diagnostic and treatment [19]. Mixed phenotype acute leukemias with myeloid and T-lymphoid features

(T/M MPAL) are defined as a separate category by the World Health Organization (WHO) classification, based on coexpression of markers such as CD3 and MPO [20]. Moreover, a subtype of T-ALL, known as early T-cell precursor leukemia (ETP-ALL), also exhibits a combination of myeloid and lymphoid surface markers [21]. Recent studies have shown that ETP-ALL and T/M MPAL are similar at the genetic and epigenetic level [22], suggesting an overlap between these two classifications. The emerging question is how CIMP leukemias, originally diagnosed as AMLs, are related to these other categories from a molecular perspective.

Here, we aimed to characterize in depth the poorly understood CIMP leukemias by integrating multiple layers of genetic and epigenetic data. Our integrated analysis revealed that these are immature leukemias with features from both AML and T-ALL, resembling ETP-ALL. We showed that hypermethylation results in repression of key lineage-specific TFs as well as reduced CTCF binding, which in turn leads to changes in chromatin architecture and secondary changes in gene expression.

RESULTS

Global DNA methylation identifies a distinct group of hypermethylated leukemias

Previous studies in separate AML cohorts independently identified clusters of patients with genome-wide hypermethylation, but no mutations typically related to DNA methylation. We jointly profiled the methylome of 16 of these patients together with 49 other primary AMLs and CD34+ cells from 3 healthy donors (Table S1). We used methyl-CpG immunoprecipitation coupled with sequencing (MCIp-seq) to assay 71,000 CpG-rich regions with an average length of 650 base pairs (bp), covering 89% of the 28,691 CpG islands in the human genome (Figure 1A). More than half of the MCIP-seq peaks are located in the proximity of a TSS, with roughly 35% of the remaining peaks within genes and 15% in intergenic regions (Figure 1B).

Principal component analysis (PCA) (Figure 1C) and hierarchical clustering (Figure 1D) revealed that CIMP leukemia constitutes a separate subgroup with strong hypermethylation, particularly at regions hypomethylated in CD34+ cells. Samples from both studied cohorts (CIMP-EMC, originally "*CEBPA*-silenced", and CIMP-UKR) clustered together, supporting the hypothesis that they belong to the same disease entity. This observation was supported by other dimensionality reduction strategies, including Uniform Manifold Approximation and Projection (UMAP) (Figures S1A-1F), as well as by hierarchical clustering of Epityper data of 190 CpG islands in 220 patients (Figure S1F). Taken together, these data confirm that CIMP leukemias are a distinct entity characterized by global hypermethylation.

The epigenetic landscape of CIMP leukemias reveals an intermediate state between T-ALL and AML

CIMP leukemias exhibit a mixed myeloid/lymphoid phenotype on the basis of their membrane markers [8]. To understand the regulatory underpinnings of this phenotype, we next compared the epigenetic and transcriptional landscape of CIMP leukemias with that of T-ALL and AML, as well as with CD34+ cells from healthy donors.

Dimensionality reduction of MCIP-seq data revealed that CIMP cases exhibit a methylation profile very close to that of most T-ALLs and markedly separate from that of AML (Figure 2A, Figure S2A). However, in terms of gene expression, H3K27ac and open chromatin, CIMP cases presented a hybrid profile between AML and T-ALL (Figures 2B-2D), and were closely related to *CEBPA* DM AML (Figures S2B-S2D) Hierarchical clustering similarly grouped CIMPs with subsets of both AML and T-ALL across all data types except for methylation (Figure S2E-S2H).

In line with their hybrid epigenetic profile and the previously reported phenotype, CIMP leukemias expressed both typical myeloid markers such as CD13, CD33, CD34 and KIT (Figure 2E) and lymphoid markers like CD7 and CD3 (Figure 2F).

In summary, CIMP leukemias appear as an intermediate entity between AML and T-ALL at the transcriptional and epigenetic levels, which explains their combination of surface markers.

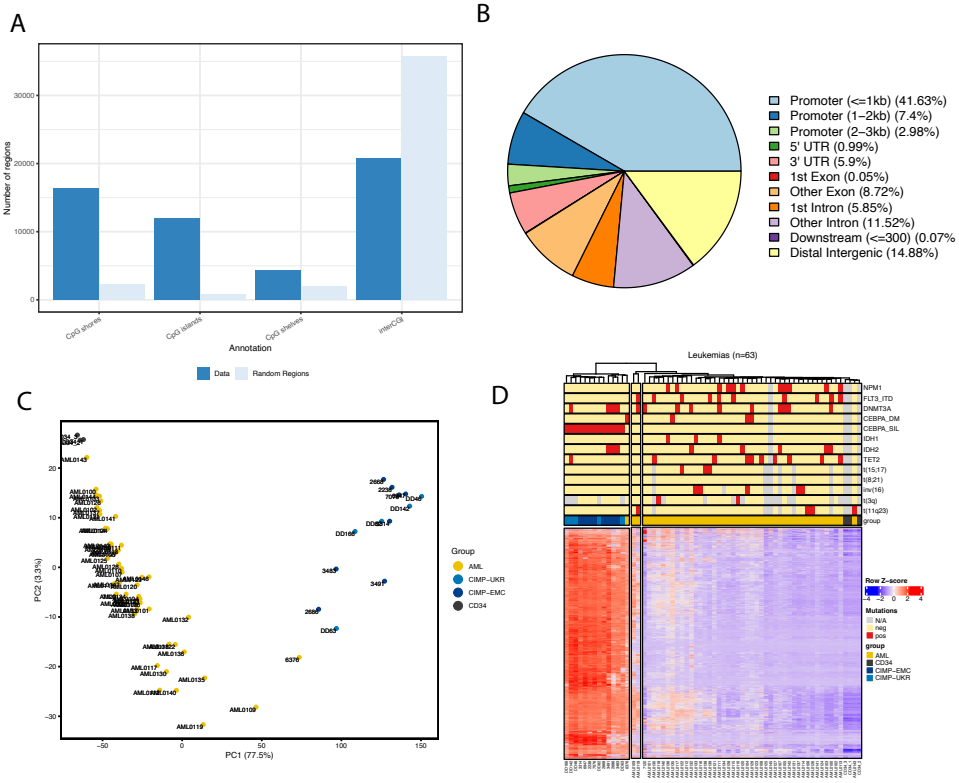
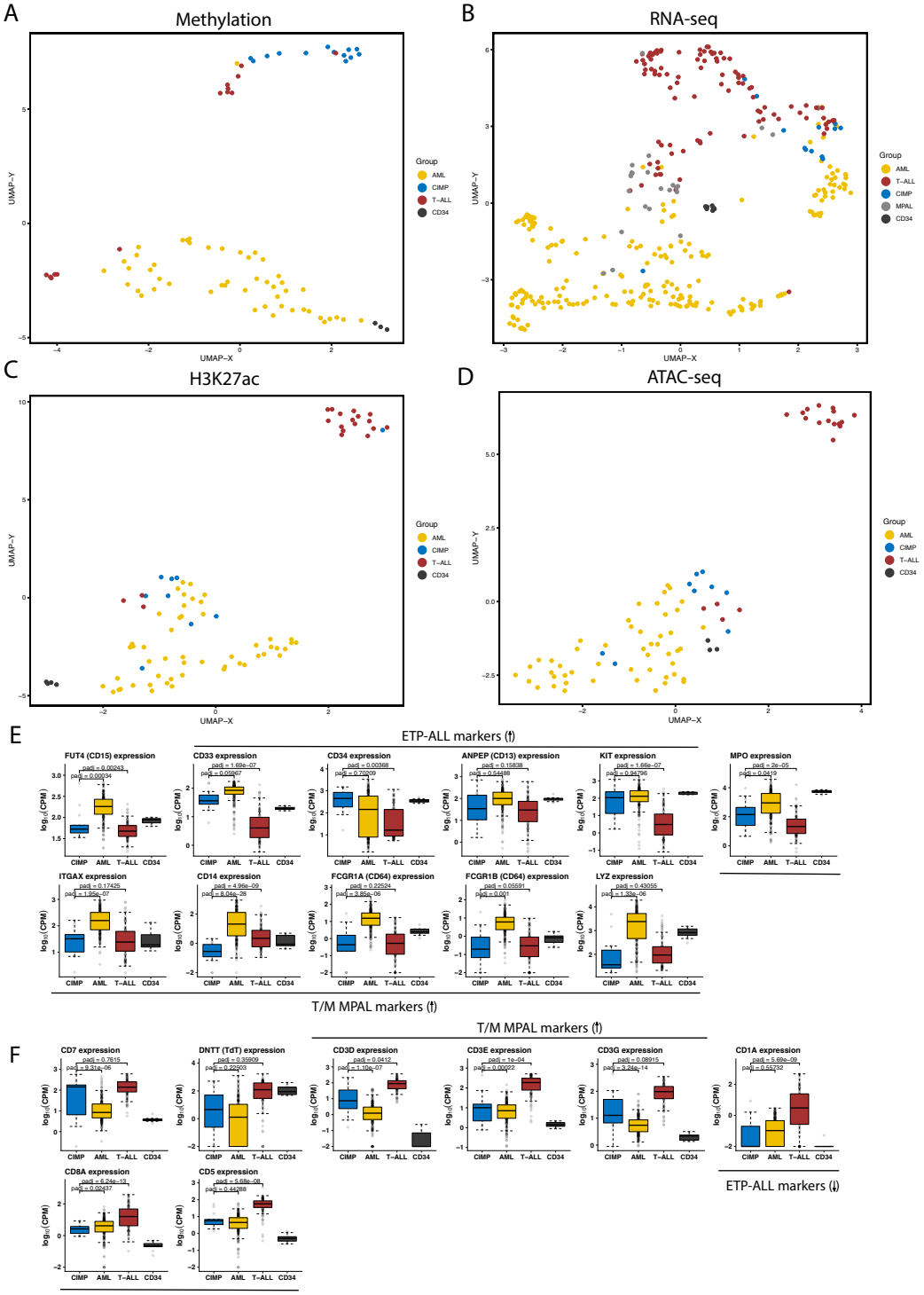


Figure 1. Methylation landscape assessed by MCIP-seq. A. Coverage of CpG islands, shores and shelves by MCIP-seq data, compared to an equal set of randomly selected regions. The plot shows enrichment at CpG-rich regions and a depletion at inter-CGI regions relative to the random set. **B.** Functional annotation of methylated regions detected by MCIP-seq. **C.** Principal component analysis (PCA) of MCIP-seq data from AML and CIMP cases. **D.** Pearson correlation heatmap of MCIP-seq data from AML and CIMP cases.

Figure 2. Epigenetic and transcriptional landscape of CIMP, AML, T-ALL and CD34+ cells. A. Dimensionality reduction with Uniform Manifold Approximation and Projection (UMAP) of methylation data measured by MCIP-seq in AML, CIMP, T-ALL and CD34+ HSPCs. **B.** Same as A, but applied to gene expression measured by RNA-seq. **C.** Same as A, but applied to histone K327 acetylation (H3K27ac) measured by ChIP-seq. **D.** Same as A, but applied to accessible chromatin measured by ATAC-seq. **E.** Expression of myeloid markers commonly used for classification in CIMP, other leukemias and healthy controls. **F.** Expression of lymphoid markers commonly used for classification in CIMP, other leukemias and healthy controls. Note: ETP-ALL is defined by absence of CD1A and CD8, weak expression of CD5 and presence of myeloid markers such as CD13, CD33, CD34 and CD177 (KIT) [119]. T/M MPAL is defined by presence of either MPO or monocytic markers (CD11c, CD14, CD64, LZE) concomitantly with CD3 expression [144].


A
Methylation
B
RNA-seq
C
H3K27ac
D
ATAC-seq
UMAP–X
UMAP–Y
Group
AML
CIMP
T–ALL
MPAL
CD34
E
ETP-ALL markers (↑)
FUT4 (CD15) expression
CD33 expression
CD34 expression
ANPEP (CD13) expression
KIT expression
MPO expression
ITGAX expression
CD14 expression
FCGR1A (CD64) expression
FCGR1B (CD64) expression
LYZ expression
T/M MPAL markers (↑)
F
T/M MPAL markers (↑)
CD7 expression
DNTT (TdT) expression
CD3D expression
CD3E expression
CD3G expression
CD1A expression
CD8A expression
CD5 expression
ETP-ALL markers (↓)
CIMP
AML
T–ALL
CD34

CIMP leukemias are genetically heterogeneous

To elucidate whether genetic aberrations lie at the base of CIMP leukemias, we conducted whole exome sequencing (WES). We found frequent (> 25% of cases) single nucleotide variants (SNVs) and indels in *NOTCH1*, *PHF6*, *MED12*, *WT1*, *IKZF1* and *JAK3*, but none of them was common to all individuals (Table S2, Figure 3A). We observed frequent copy number alterations (CNAs) compared to a panel of CD34+ controls (Tables S3-S4), albeit none of them were present in more than 3 patients (Figure 3B, Supplementary Figures 3A-3B). Recurrent focal CNAs were identified in a number of genes related to leukemia and hematopoiesis, among which deletions of a region containing *NF1*, *EVI2A* and *EVI12B* were particularly frequent (n=6) (Figure 3C). Interestingly, 3 of the patients who carried NF1 deletions also exhibited point mutations in that gene, presumably in the other allele.

RNA-seq data did not reveal any recurrent fusion genes, although a few patients carried fusions previously reported in leukemia (Figure 3D, Table S5). Of note, 6/13 patients analyzed with RNA-seq did not harbor any fusion gene.

Altogether, CIMP leukemias constitute a genetically heterogeneous subgroup, defined by epigenetic rather than genetic commonalities. As a whole, their mutational profiles are comparable to those of other acute leukemias of ambiguous lineage, especially ETP-ALL (Supplementary Results), suggesting significant overlap between these entities.

Transcriptional signatures suggest similarity to ETP-ALL with an early lymphoid-biased progenitor as the cell of origin

To investigate lineage relationships at the transcriptional level, we conducted gene set enrichment analysis (GSEA) (Figure 4A, Tables S9-12, Supplementary Figures 4A-D). In line with the mixed phenotype of these leukemias, myeloid gene sets (e.g. *Ebert_Myeloid.Up500*) were downregulated when compared to AML, but upregulated when compared to T-ALL; the reverse was true for T-lymphoid genes. Interestingly, the top results from the comparison with T-ALL were gene sets derived from ETP-ALL relative to other T-ALLs (e.g. *ETP-ALL_Zhang_Up*), as well as gene sets related to hematopioietic stem cells (HSCs) (e.g. *Dick_HSC250*). A comparison with CD34+ cells revealed upregulation of T-cell signatures, including ETP, but downregulation of HSC signatures.

In a single sample GSEA with a selected number of hematopoietic-related gene sets, the CIMP group exhibited enrichment for HSC genes as well as myeloid and lymphoid signatures halfway between AML and T-ALL (Figure 4B, Supplementary Figures 4E-H). Similarly, analysis of transcriptional signatures of cell types derived from publicly available data [23,24] using CIBERSORTx [25] showed enrichment for cycling HSCs, immature neutrophils, GMPs and CLPs, depending on the selected signature matrix (Figure 4C, Supplementary Figure 4I).

Altogether, these results emphasize the similarities between CIMP and ETP-ALL and suggest that the cell of origin in CIMP leukemias is an early progenitor committed to the lymphoid lineage.

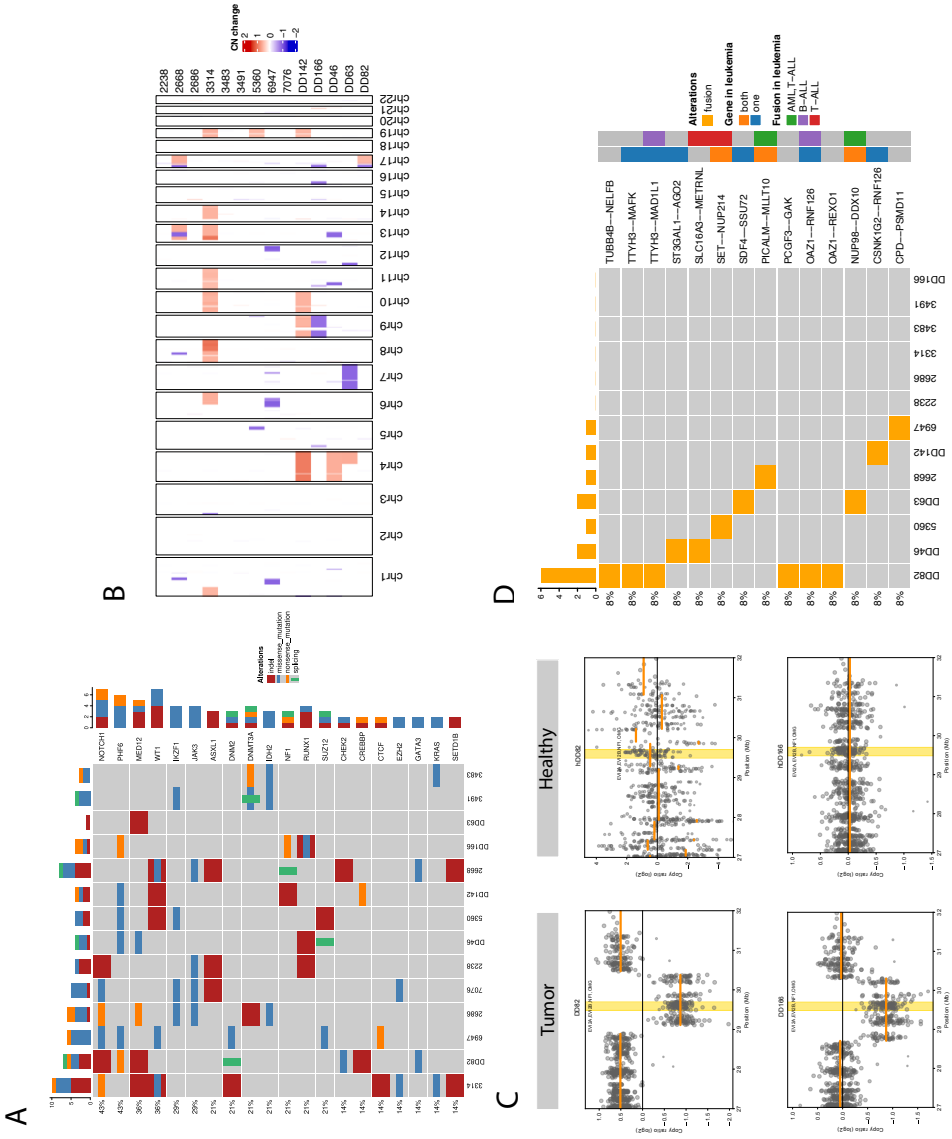


Figure 3. Mutational landscape of CIMP leukemias. A. Oncoprint displaying single nucleotide variants (SNVs) and small inserts and deletions found in genes mutated in at least 2% of the cohort (N=14). Columns correspond to patients and rows correspond to genes, ranked by mutational frequency. Variant calling was performed with an ensemble of tools on whole exome sequencing (WES) data. Different variants are indicated in different colors as shown in the legend of the plot. **B.** Heatmap of copy number alterations (CNAs) in CIMP cases, detected using CNVkit on WES data. Red indicates copy number gains (CNG) and blue indicates copy number losses (CNL). **C.** Scatter plot showing copy number ratios (grey dots) and segmentation calls (orange lines) in the *EVI2A/B*,*NF1* locus, for two CIMP cases where both tumor and healthy samples are available. **D.** Oncoprint displaying fusion genes detected by at least 3 different software tools and not commonly found in healthy individuals. Annotations indicate which fusion genes have been reported in different leukemia sequencing projects [112–115] or fusion genes whose interacting partners are involved in leukemia according to the Disgenet database [116].

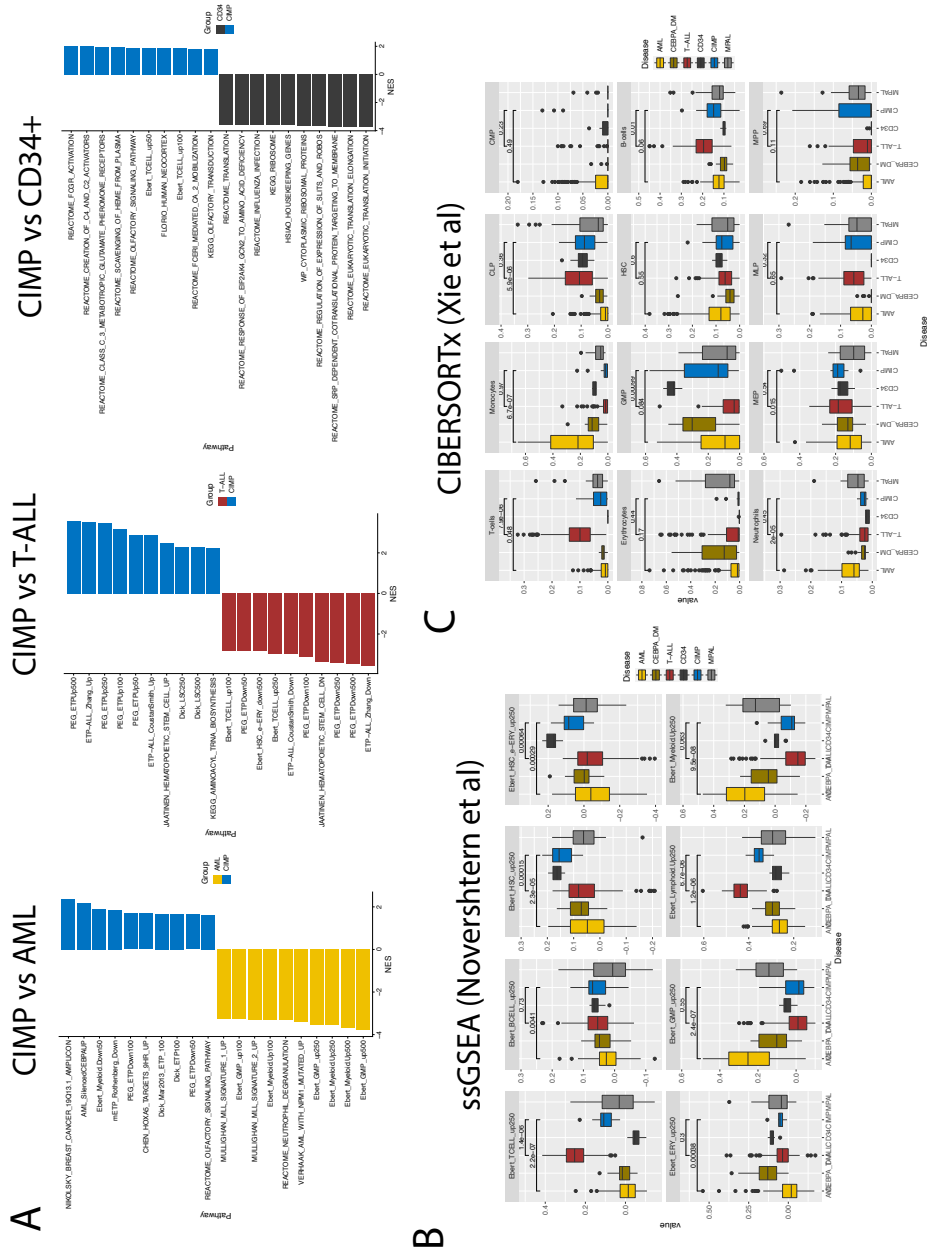


Figure 4. Transcriptional signatures of CIMP and other leukemias. A. Bar plot showing the top results from gene set enrichment analysis (GSEA) conducted on a custom version of the MSigDB C2 collection. The analysis was conducted on differentially expressed genes in CIMP relative to AML (left panel), T-ALL (middle) and CD34+ HSPCs (right). **B.** Box plot displaying the scores of a single sample GSEA (ssGSEA) analysis with gene sets derived from various hematopoietic fractions [145]. The analysis was conducted on fragments per million (FPM)-normalized RNA-seq data from multiple leukemia subgroups. **C.** Box plot displaying the CIBERSORTx scores using a signature matrix derived from the Atlas of Human Blood Cells [24]. The analysis was performed on a mixture matrix containing RNA-seq raw counts from multiple leukemia subgroups.

Promoter methylation changes lead to silencing of critical hematopoietic factors in CIMP leukemias

Next, we investigated the distribution and effects of methylation in CIMPs in relation to other leukemias and normal controls. CIMPs exhibited profound hypermethylation at CGIs, promoters, enhancers, transcriptional start sites (TSS) and gene bodies when compared to AML or HSPCs. On the other hand, methylation levels were similar in T-ALL (Figure 5A, Figure S5A-B).

Promoter methylation levels were the highest in CIMP, followed by T-ALL, AML and HSPCs (Figure 5A, 5B). These differences are consistent with the higher levels of methylation in the lymphoid lineage [26–28]. However, hypermethylation was not present in terminally differentiated cells of any of those lineages (Figure S5C, S5D). Differential methylation analysis confirmed extensive hypermethylation in CIMP compared to AML (CIMP vs AML), and to T-ALL to a lesser extent (Figure 5D-5E, Table S13). Hypermethylation was more pronounced in regions marked by both H3K4me3 and H3K27me3 (Figure 5C, Figure S5E-S5F), typically referred to as "bivalent promoters" [29]. This is in line with previous reports showing that bivalent promoters are more susceptible to DNA hypermethylation in both cancer cell lines and primary tumors [30]. Furthermore, gene set enrichment analysis (GSEA) of genes in the vicinity of DMRs confirmed preferential hypermethylation of H3K27me3 targets in CIMP relative to AML, T-ALL and CD34+ HSPCs (Figure 5F, Figures S5C-E).

Some of the differentially methylated regions (DMR) with strongest increases were adjacent to genes such as *SPRED1*, *LEF1*, *PLK2*, *MEIS1* or *TLE4*, with a known involvement in either leukemia or hematopoiesis (Figure 5E, Figure S5G). Indeed, GSEA revealed an enrichment of transcription factor (TFs) genes in methylated regions relative to both AML and CD34+, as well as genes involved in cell commitment (Figure 5G). Thus, we next analyzed methylation at promoter regions defined by the FANTOM consortium. Of the 14321 promoters analyzed, 2644 corresponded to known TFs, belonging to 1427 unique genes. In the CIMP cohort, 19.6% of the 2644 TF promoters were hypermethylated with respect to AML alone (FDR < 0.05), 8.0% with respect to both AML and T-ALL and only 0.5% relative to T-ALL alone (Figure 6A, Table S14). 1.3% and 3.0% of TF promoters exhibited higher methylation in T-ALL and AML, respectively, than in CIMP. As expected, integration of gene expression data confirmed that CIMP vs AML hypermethylation was accompanied by widespread gene silencing, with methylation levels negatively correlating with gene expression (Figure S6A, Figure S6B). A total of 101 TFs with silenced promoters were downregulated, including several hematopoietic regulators and genes known to be involved in leukemia, such as *CEBPA, HOXB9*, *CEBPD*, *MECOM*, *IRF4* and *KLF2* (Figure 6B-C, Figure S6C). Among the few TF genes differentially methylated between CIMP and T-ALL was *LEF1*, which participates in early stages of thymocyte maturation [31] and is also crucial for neutrophilic granulopoiesis [32].

In summary, many critical TF are silenced by methylation in CIMP leukemias, which possibly explains the intermediate epigenetic state of these leukemias, as well as their differentiation arrest.

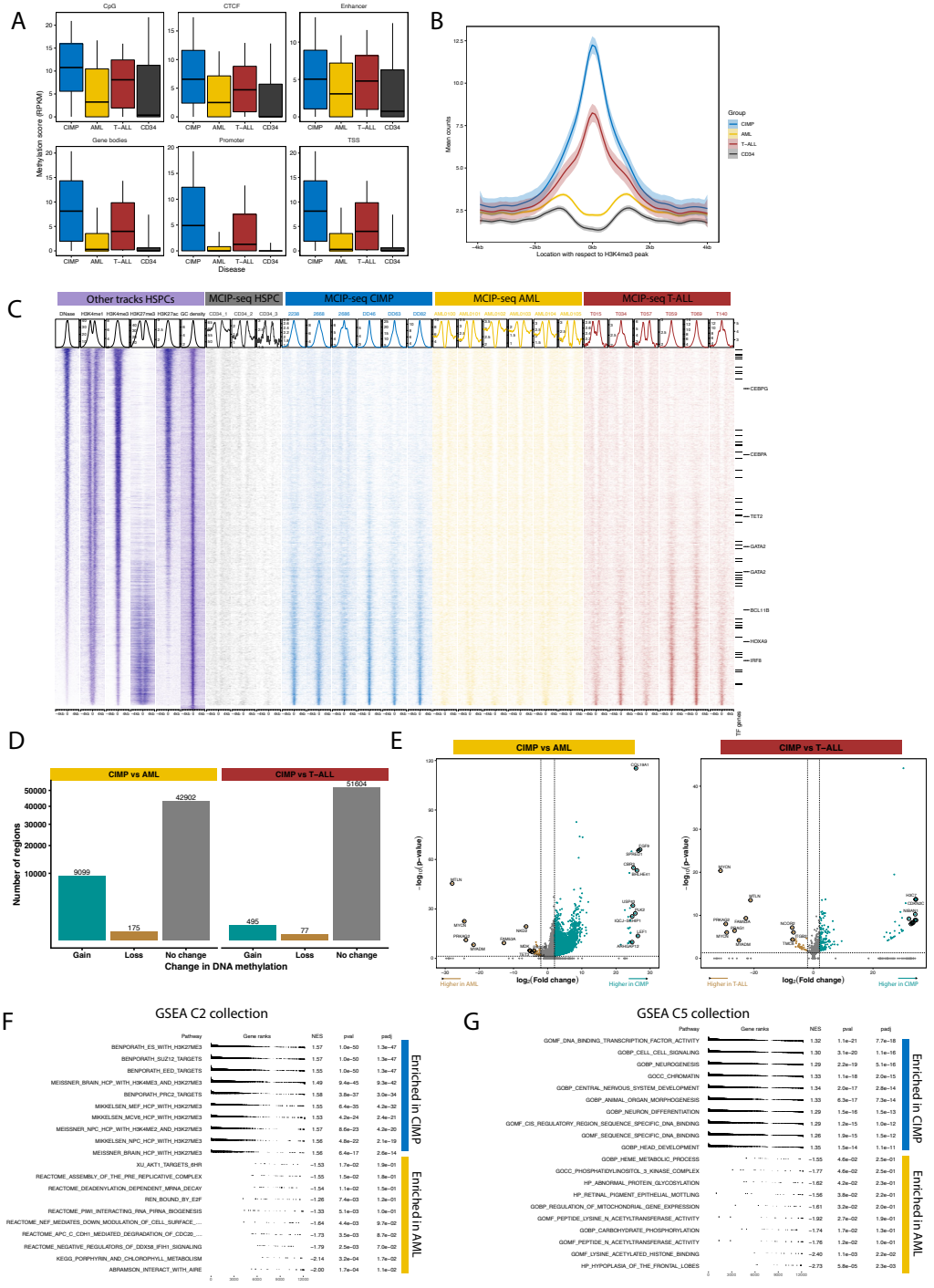

A
CpG
CTCF
Enhancer
Gene bodies
Promoter
TSS
Methylation score (RPKM)
CIMP
AML
T−ALL
CD34
Disease
B
Mean counts
Group
Location with respect to H3K4me3 peak
C
Other tracks HSPCs
MCIP-seq HSPC
MCIP-seq CIMP
MCIP-seq AML
MCIP-seq T-ALL
CEBPG
CEBPA
TET2
GATA2
BCL11B
HOXA9
IRF8
TF genes
D
CIMP vs AML
CIMP vs T−ALL
Number of regions
9099
175
42902
495
77
51604
Gain
Loss
No change
Change in DNA methylation
E
CIMP vs AML
CIMP vs T−ALL
−log10(p-value)
log2(Fold change)
Higher in AML
Higher in CIMP
Higher in T-ALL
F
GSEA C2 collection
Pathway
Gene ranks
NES
pval
padj
BENPORATH_ES_WITH_H3K27ME3
BENPORATH_SUZ12_TARGETS
BENPORATH_EED_TARGETS
MEISSNER_BRAIN_HCP_WITH_H3K4ME3_AND_H3K27ME3
BENPORATH_PRC2_TARGETS
MIKKELSEN_MEF_HCP_WITH_H3K27ME3
MIKKELSEN_MCV8_HCP_WITH_H3K27ME3
MEISSNER_NPC_HCP_WITH_H3K4ME2_AND_H3K27ME3
MIKKELSEN_NPC_HCP_WITH_H3K27ME3
MEISSNER_BRAIN_HCP_WITH_H3K27ME3
XU_AKT1_TARGETS_6HR
REACTOME_ASSEMBLY_OF_THE_PRE_REPLICATIVE_COMPLEX
REACTOME_DEADENYLATION_DEPENDENT_MRNA_DECAY
REN_BOUND_BY_E2F
REACTOME_PIWI_INTERACTING_RNA_PIRNA_BIOGENESIS
REACTOME_NEF_MEDIATES_DOWN_MODULATION_OF_CELL_SURFACE_...
REACTOME_APC_C_CDH1_MEDIATED_DEGRADATION_OF_CDC20_...
REACTOME_NEGATIVE_REGULATORS_OF_DDX58_IFIH1_SIGNALING
KEGG_PORPHYRIN_AND_CHLOROPHYLL_METABOLISM
ABRAMSON_INTERACT_WITH_AIRE
Enriched in CIMP
Enriched in AML
G
GSEA C5 collection
GOMF_DNA_BINDING_TRANSCRIPTION_FACTOR_ACTIVITY
GOBP_CELL_CELL_SIGNALING
GOBP_NEUROGENESIS
GOCC_CHROMATIN
GOBP_CENTRAL_NERVOUS_SYSTEM_DEVELOPMENT
GOBP_ANIMAL_ORGAN_MORPHOGENESIS
GOBP_NEURON_DIFFERENTIATION
GOMF_CIS_REGULATORY_REGION_SEQUENCE_SPECIFIC_DNA_BINDING
GOMF_SEQUENCE_SPECIFIC_DNA_BINDING
GOBP_HEAD_DEVELOPMENT
GOBP_HEME_METABOLIC_PROCESS
GOCC_PHOSPHATIDYLINOSITOL_3_KINASE_COMPLEX
HP_ABNORMAL_PROTEIN_GLYCOSYLATION
HP_RETINAL_PIGMENT_EPITHELIAL_MOTTLING
GOBP_REGULATION_OF_MITOCHONDRIAL_GENE_EXPRESSION
GOMF_PEPTIDE_LYSINE_N_ACETYLTRANSFERASE_ACTIVITY
GOBP_CARBOHYDRATE_PHOSPHORYLATION
GOMF_PEPTIDE_N_ACETYLTRANSFERASE_ACTIVITY
GOMF_LYSINE_ACETYLATED_HISTONE_BINDING
HP_HYPOPLASIA_OF_THE_FRONTAL_LOBES
Enriched in CIMP
Enriched in AML

Figure 5. Functional assessment of methylation differences between CIMP and other leukemias. A. Box plot showing methylation levels of CIMP, AML, T-ALL and CD34+ cells at different genomic features derived from in-house data and publid databases. **B.** Average methylation levels (MCIP-seq) of different leukemias and healthy cells at putative promoter regions, defined as 4-kb regions surrounding the center of H3K4me3 ChIP-seq peaks in CD34+ HSPCs. **C.** Tornado plot depicting methylation (MCIP-seq) at putative HSPC promoters, sorted by chromatin accessibility in HSPCs. The color code distinguishes different types of leukemia and HSPCs, and the intensity reflects the degree of methylation. The HSPC tracks in purple were downloaded from ENCODE [146]and show chromatin accessibility (DNase-seq) as well as histone marks for enhancers (H3K4me1), promoters (H3K4me3), activation (H3K27ac) and repression (H3K27me3). GC density was downloaded from the UCSC browser [147]. **D.** Bar plot of differentially methylated regions (DMR) in supervised comparisons of MCIP-seq peaks between CIMP and AML (left) or T-ALL (right). A threshold of FDR < 0.05 and |log2 FC| > 1 was used to determine significant DMRs. NS = not significant. **E.** Volcano plot of DMRs annotated with the closest genes in the linear genome. Regions with a FDR < 0.05 and log2 fold change > 2 are highlighted. **F.** Summary of results of pre-ranked GSEA conducted on genes in the vicinity of DMRs between CIMP and AML, using the FDR as the ranking value. The C2 (left) and C5 (right) MSigDB collections were used in the analysis.

Loss of *CEBPA* plays a critical role in shaping the leukemic epigenome

Among the TFs downregulated by methylation was *CEBPA*, the loss of which was originally identified as the defining feature of the CIMP-EMC cohort [8], also known as *CEBPA*-silenced. In line with the initial reports, a recurrent observation across the analyses of epigenomics data was that CIMP leukemias exhibited a profound similarity with double mutant AML (Figure 2A-D, Figure S2A-D). Double *CEBPA* mutations define an AML subtype (CEBPA DM) with a distinct gene expression profile, comparable to that of CIMP leukemias[8,33]. These patients typically exhibit a combination of N- and C-terminal mutations in the CEBPA protein that disrupt its normal function [34]. Moreover, CIMPs also clustered in the vicinity of AMLs with t(8;21), a chromosomal aberration that produces a RUNX1-RUNXT1 fusion protein, which inhibits the expression of *CEBPA* [35]. The similarity between epigenetic profiles of *CEBPA* DM AMLs and CIMP suggests that loss of function of *CEBPA*, either by genetic or epigenetic hits, drives the acquisition of a distinct epigenetic and transcriptional landscape.

To further investigate this possibility, we used ChromVAR to estimate the activity of TFs based on deviations of chromatin accessibility measured ATAC-seq data (details in Supplementary Results). The C/EBP family of TFs was among the top 30 with the largest variability across the whole cohort (Figure S10C). Importantly, they were among the few TFs with a significant loss of activity in CIMP relative to AML, whereas they displayed the largest increases of accessibility in AML compared to T-ALL (Figure S10D). This underscores the importance of *CEBPA* as a critical determinant of cell identity and supports the notion that its loss in CIMP leukemias underlies their unique differentiation status.

Genome-wide hypermethylation leads to widespread loss of CTCF binding

Since DNA methylation may weaken the binding of CTCF [36,37], the hypermethylation observed at CTCF binding sites (Figure 5A) suggested a possible loss of CTCF binding at those locations. Indeed, CTCF ChIP-seq (Figures 7A-B) showed that global CTCF levels were lower in CIMP than in AML and T-ALL (Figures 7C-D). A supervised analysis confirmed widespread loss of CTCF binding in CIMP with respect to AML, and to a lesser extent compared to T-ALL (Figure 7E, Figure S7A, Table S15).

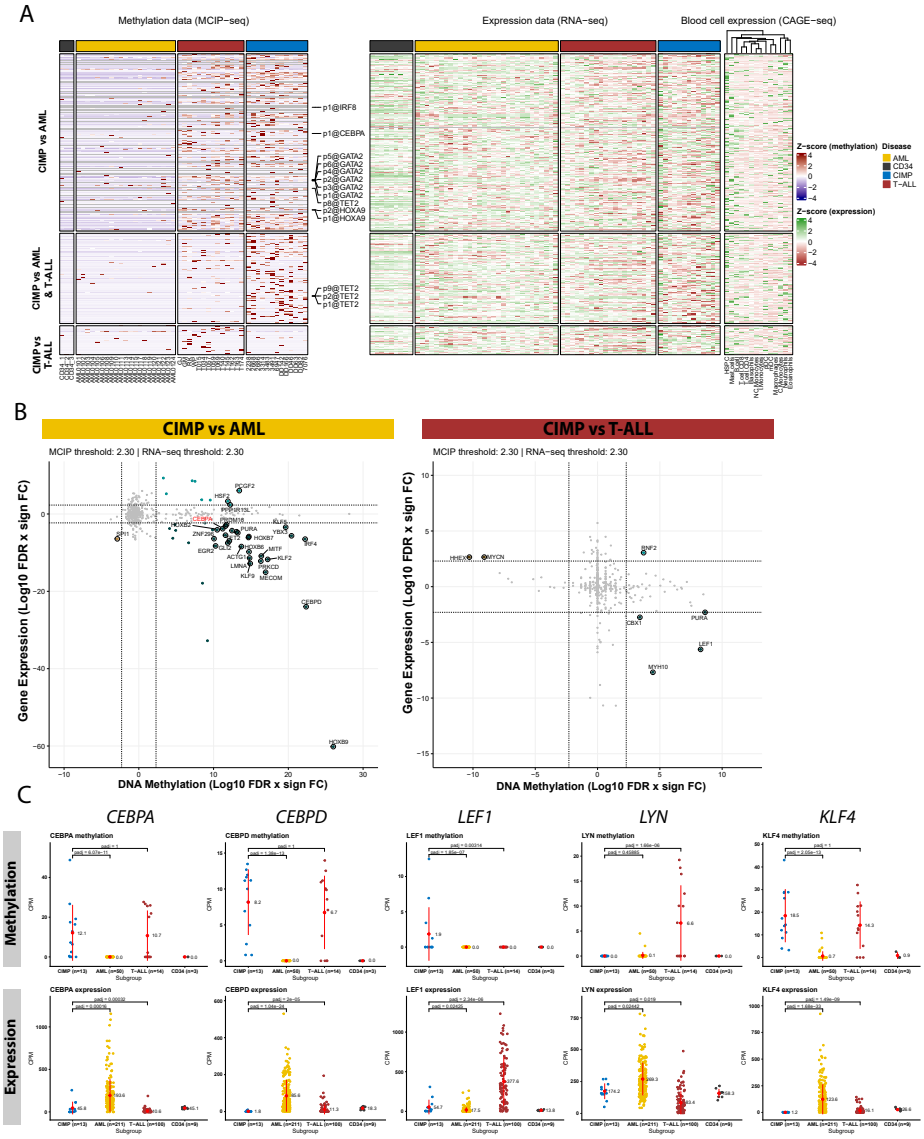


Figure 6. Integration of methylation and gene expression data reveals silencing of transcription factors involved in hematopoiesis. A. Heatmap displaying normalized methylation levels (MCIP-seq) at promoters of differentially methylated transcription factors in CIMP with respect to AML (CIMP vs AML), T-ALL (CIMP vs T-ALL) or both (CIMP vs AML & T-ALL), including both hyper- and hypomethylated regions. The heatmap in the middle shows normalized expression levels (RNA-seq) of the same genes in leukemia cells and healthy HSPCs. The rightmost heatmap presents normalized gene expression (CAGE-seq) in different healthy cells. **B.** Starburst plot depicting changes in gene expression (Y axis) and methylation (X axis) between CIMP and AML (left) and T-ALL (right). The values are the log10 of the false discovery rate (FDR) with the sign of the fold change in comparisons by DESeq2; regions with a FDR < 0.05 and log2 fold change > 2 are highlighted. Only transcription factors involved in hematopoiesis (GO term GO:0030097), cancer (COSMIC database [135]) or leukemia (Disgenet [116]) are shown, among which the top 30 genes by log2 fold change are annotated. **C.** Jitter plots showing methylation (top) and expression (bottom) of a few selected genes in CIMP and other leukemias, as well as HSPCs.

For additional insight on the interplay between methylation and CTCF binding, we integrated these data with MCIP-seq, which covered around 25% of the CTCF sites detected by ChIP-seq (Figure S7B, Table S16). The CTCF levels genome-wide were higher in regions with low DNA methylation (Figure 7D). Accordingly, gain of DNA methylation in CIMP cases correlated with loss of CTCF binding in the same regions when compared to AML (rho = -0.27, p-value = 3.04×10^{-266}) (Figure 7F, Figure S7C, Table S17). No meaningful correlation was observed when comparing CIMP and T-ALL, possibly due to the small differences in methylation between the two groups. Conversely, increases in methylation in CIMP relative to AML were higher at CTCF binding sites (Figure S7D), especially at those that were lost (Figure 7G), suggesting those are particularly prone to methylation changes.

The invariability of CTCF binding at some regions (Figure S7E-S7F) is in keeping with previous studies indicating that only certain CTCF binding sites are sensitive to methylation, such as the ones with CpG in their motif [38,39]. To explore this possibility, we computed frequency of CpG dinucleotides at every position of the canonical CTCF motif, which exhibits two peaks at position 5 and 15 respectively (Figure S7H). CTCF motifs found in regions with loss of CTCF binding and hypermethylation exhibited CpGs at those two positions more frequently than regions where CTCF binding remained unchanged or increased (Figure S7G, S7I).

Genome-wide hypermethylation is accompanied by changes in 3D organization

Given the prominent role of CTCF in the stabilization of cohesin-mediated chromatin loops [40,41], we conducted in situ Hi-C experiments on CIMP (n=9), AML (n=5), T-ALL (n=4) and HSPCs (n=3) to assess changes in 3D genome organization. Detection of 3D organization features was conducted in aggregate, yielding a total of 4537 TADs and 9443 loops across all datasets. Roughly 40-50% of the called CTCF sites overlapped with TAD boundaries and 10-20% with loop anchors, with minimal differences between variable and unchanged peaks (Figure S8A).

We detected a clear separation between AML and other leukemias both at the level of TADs (Figure 8A, Figure S8B, Figure S8D) and loops (Figure 8B, Figure S8C, Figure S8E). Most CIMP cases clustered together with T-ALLs, with a few (DD46, DD63) exhibiting stronger similarity with AML. CD34+ cells were excluded from these analyses because they were dramatically different from all other samples, masking smaller differences between other groups (Figures S8F-G). Supervised comparisons of differential loops or interactions (DIs) and variable TADs (ΔTADs) confirmed that differences between CIMP and AML were larger than between CIMP and T-ALL, but somewhat smaller than between AML and T-ALL (Figures 8C-D).

Contrary to our expectations, we did not observe a widespread depletion of chromatin loops or TADs upon loss of CTCF binding in CIMP cases. However, 72% of the loops lost in CIMP relative to AML exhibited decreased CTCF binding in at least one of their anchors, compared

to 59% in gained interactions (Figure 8E). Moreover, the average decrease in CTCF binding was significantly higher in lost interactions (Figure S8I). Therefore, while most changes of chromatin conformation in CIMP seem to occur independently of hypermethylation-derived loss of CTCF binding, the latter has a contributing role.

Next, we conducted an unbiased survey of ΔTADs (Tables S18-S19) and DIs (Tables S20-S21) with associated changes in CTCF binding and potential implications for gene expression. When comparing CIMP and AML, we found 61 ΔTADs containing differentially expressed genes with loss of CTCF binding at their boundaries and 71 differential enhancer-promoter loops, whose interaction strength strongly correlated with the expression of genes they contacted (rho = 0.67, $p = 1.8 \times 10^{-10}$, Figure 8F). Among others, loss of insulation was detected at the TADs containing *KLF4* (Figure 8G) and *CEBPD* (Figure 8H), both of which also displayed decreased chromatin interactions, which was accompanied by reduced CTCF binding. Interestingly, their promoters were also methylated, suggesting a possible cooperation between distinct epigenetic mechanisms in repression. Examples of gained enhancer-promoter interactions included a loop connecting *GATA3* with a nearby enhancer element that is specific to CIMP (Figure 8I) and a loop involving the promoter of *DNMT3B* (Figure S8M). More details are provided in the Supplementary Results.

In sum, CIMPs exhibit partial rewiring of chromatin interactions when compared to AML, of which only a fraction are attributable to loss of CTCF. However, this mild remodeling results in the misexpression of some essential TFs. Very few 3D genome differences could be detected between CIMP and T-ALL, in line with the notion that these leukemias originate from a lymphoid-biased cell.

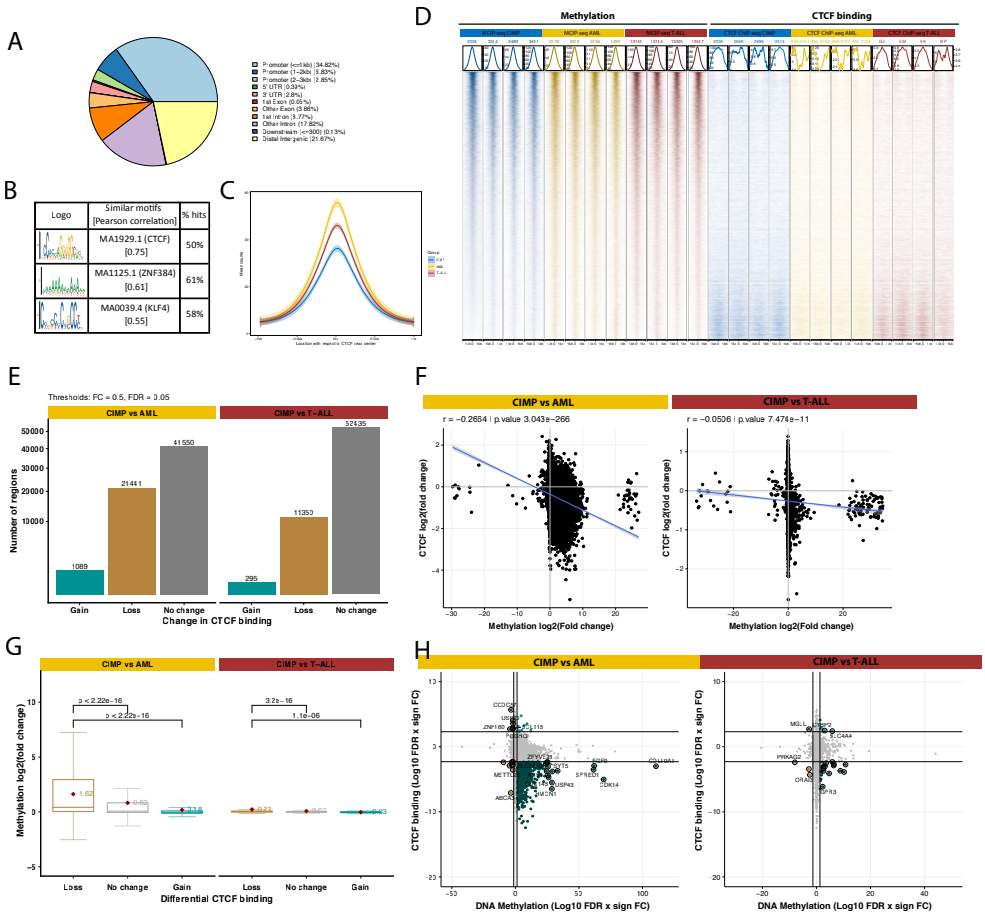


Figure 7. Hypermethylation in CIMP leukemias leads to loss of CTCF binding. A. Functional annotation of aggregated CTCF binding peaks detected by ChIP-seq **B.** Top 3 results of *de novo* motif analysis in CTCF peaks with rGADEM. The second column indicates similar motifs in the JASPAR database (v2020) based on Pearson correlation and the third column shows the percentage of peaks harboring the motif. **C.** Average CTCF binding of different leukemias in 1-kb regions surrounding the center of CTCF ChIP-seq peaks on a consensus master list. **D.** Tornado plots depicting methylation levels and CTCF binding at the 25,000 most variable CTCF peaks found in at least 4 patients of the entire cohort. Four representative samples of each leukemia type (CIMP, AML and T-ALL) are presented. The plot above shows the average signal around the center of the peaks for each patient. An inverse correlation between methylation and CTCF binding can be observed. **E.** Bar plot of differentially methylated regions (DMR) in supervised comparisons of MCIP-seq peaks between CIMP and AML (left) or T-ALL (right). A threshold of FDR < 0.05 and |log2 FC| > 1 was used to determine significant DMRs. NS = not significant. **F.** Scatter plot showing the inverse correlation between differences in promoter methylation (X axis) and differences in CTCF binding (Y axis) in CIMP compared to AML (left panel) and to T-ALL (right panel). The values correspond to the log10 of the FDR with the sign of the fold change. **G.** Box plot displaying methylation changes in relation to differences in CTCF binding between CIMP and AML (left) or T-ALL (right). **H.** Starbust plot depicting changes in gene expression (Y axis) and methylation (X axis) between CIMP and AML (left) and T-ALL (right). The values are the log10 of the false discovery rate (FDR) with the sign of the fold change; regions with a FDR < 0.05 and log2 fold change > 2 are highlighted. The closest gene is annotated for the top 30 regions with the largest differences.

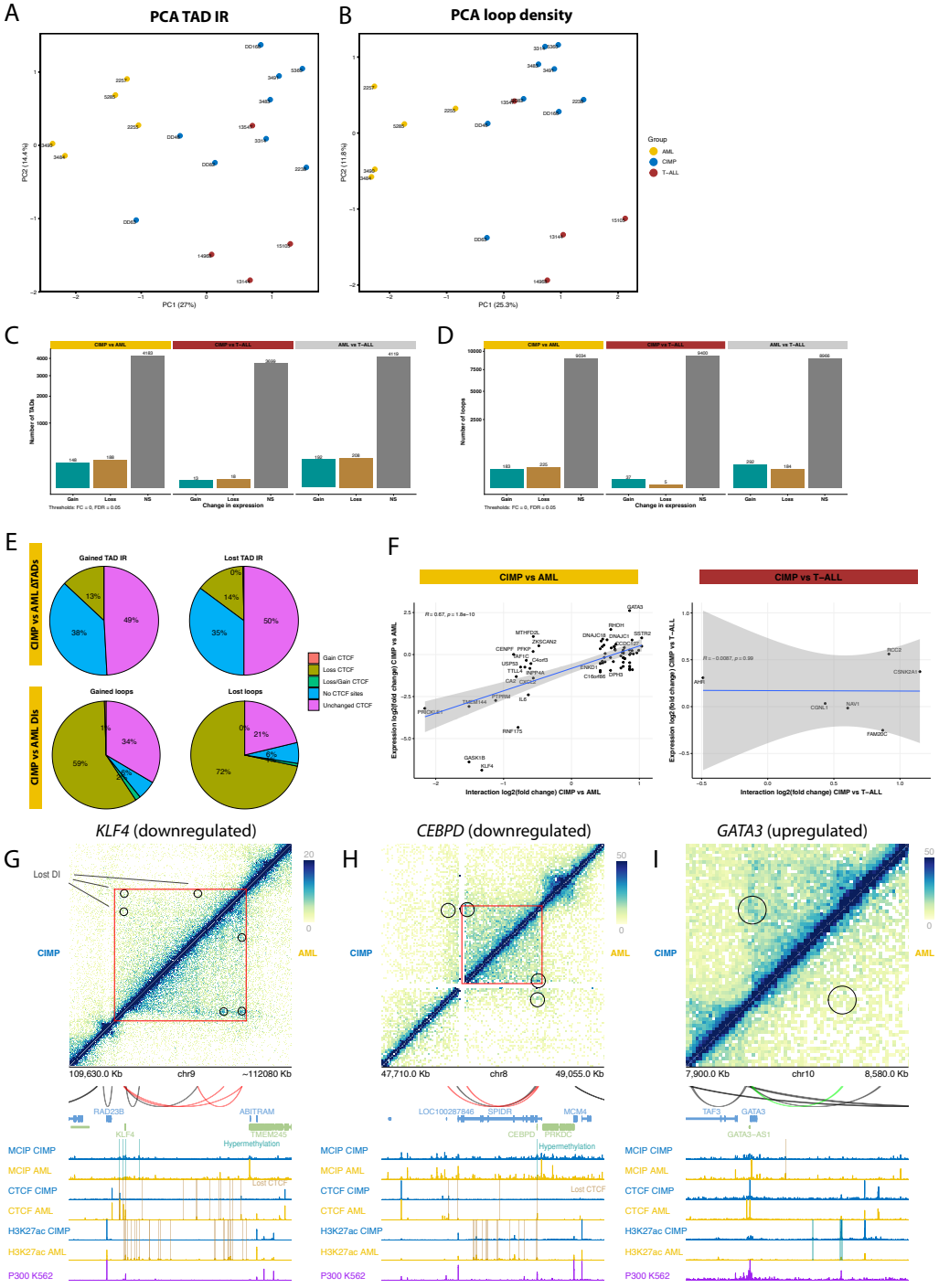

A
PCA TAD IR
PC1 (27%)
PC2 (14.4%)
B
PCA loop density
PC1 (25.3%)
PC2 (11.8%)
Group
AML
CIMP
T−ALL
C
CIMP vs AML
CIMP vs T−ALL
AML vs T−ALL
Number of TADs
Gain
Loss
NS
Change in expression
D
Number of loops
E
CIMP vs AML dTADs
CIMP vs AML DIs
Gained TAD IR
Lost TAD IR
Gained loops
Lost loops
Gain CTCF
Loss CTCF
Loss/Gain CTCF
No CTCF sites
Unchanged CTCF
F
CIMP vs AML
CIMP vs T−ALL
Expression log2(fold change) CIMP vs AML
Interaction log2(fold change) CIMP vs AML
Expression log2(fold change) CIMP vs T−ALL
Interaction log2(fold change) CIMP vs T−ALL
KLF4 (downregulated)
CEBPD (downregulated)
GATA3 (upregulated)
G
H
I
Lost DI
CIMP
AML
109,630.0 Kb
chr9
~112080 Kb
47,710.0 Kb
chr8
49,055.0 Kb
7,900.0 Kb
chr10
8,580.0 Kb
RAD23B
ABITRAM
KLF4
TMEM245
Hypermethylation
LOC100287846
SPIDR
MCM4
CEBPD
PRKDC
TAF3
GATA3
GATA3−AS1
Lost CTCF
MCIP CIMP
MCIP AML
CTCF CIMP
CTCF AML
H3K27ac CIMP
H3K27ac AML
P300 K562

Figure 8. Chromatin interaction landscape of CIMP and other leukemias. A. PCA plot of TAD inclusion ratios (IR) calculated by HOMER in Hi-C data from CIMP, AML and T-ALL. **B.** PCA plot of loop density scores calculated by HOMER in Hi-C data. **C.** TADs with differential IRs between different leukemias as calculated by DESeq2. Only TADs with a log2 fold change larger than 0 and FDR < 0.05 have been considered. **D.** Same as C, but for loop density scores. **E.** Distribution of gains or losses in CTCF binding in variable TADs (top) or differential interaction (bottom) when comparing CIMP vs AML. Lost DIs are enriched for sites with decreased CTCF binding. **F.** Correlation between changes in interaction strength and expression levels of the genes contacted by those loops. **G.** Aggregated HIC heatmap of the *KLF4* locus, comparing interactions between the CIMP (uppermost triangle, n=5) and AML groups (bottom triangle, n=5). ΔTADs are marked with a triangle in each half and DIs are indicated with circles. Underneath, loops detected in this region are shown in black, if they are invariable across conditions, and in red or blue if they are gained or lost in CIMP relative to AML, respectively. The tracks below display MCIP-seq, CTCF ChIP-seq and H3K27ac ChIP-seq from CIMP (n=4) and AML (n=4). Peaks gained in CIMP are highlighted in turquoise, whereas lost peaks are highlighted in light brown. The last track shows p300 binding measured by ChIP-seq in the K562 cell line. **H.** Same as G., but the *CEBPD* locus is shown instead. **I.** Same as G., but the *GATA3* locus is shown.

DISCUSSION

In this study, we investigated a group of leukemias with a CpG Island Methylator Phenotype (CIMP) and mixed myeloid/lymphoid features, which possibly result from a hybrid epigenetic landscape. The methylation patterns of these leukemias are very similar to those of a large fraction of T-ALL cases. This is in contrast with our previous report [9], a discrepancy that can be attributed to the fact it compared CIMPs and T-ALLs in isolation. Indeed, minor differences in methylation between the two groups segregate them when AMLs are excluded from the analysis (Figure S11A). On the other hand, clustering of expression data situates CIMPs between AML and T-ALL, whereas both H3K27ac ChIP-seq and ATAC-seq data suggest more proximity to AML. Altogether, these results indicate that CIMPs harbor an intermediate epigenetic state between lymphoid and myeloid leukemias. In the absence of common genetic lesions, this shared epigenetic profile seems to be the main defining feature of CIMP leukemias.

The existence of acute leukemias with a mixed myeloid/lymphoid phenotype has long been recognized [42]. The 2022 WHO classification identifies T/M MPALs based on a reduced number of immunophenotypic markers, but the CIMP cases identified here do not always conform to these criteria (Figure 2E-F). Moreover, the mutational profile of MPALs does not exactly match our findings [20]. Ultimately, T/M MPAL is a broad category that may partially overlap with CIMP, but possibly encompasses multiple subtypes of leukemia with variable cells of origin and pathogenic mechanisms. Another well-known entity with ambiguous lineage is ETP-ALL [21], originally defined by a gene expression signature derived from murine ETPs, but typically identified by associated membrane markers. The comparable expression of those markers in CIMP cases suggests both are similar entities, a notion further supported by their comparable mutational profiles and ETP-ALL signatures in GSEA. The distinction between CIMP and ETP-ALL may stem from their original definition as a subtype of a different disease (AML and T-ALL, respectively) and with a focus on different biological mechanisms.

The hybrid epigenetic state of these leukemias and their mixed phenotype invites the question of what their cell of origin is. DNA methylation is a stable mark of epigenetic memory that maintains cell identity across cell divisions [43] , which has been exploited to predict cell types [27] and identify the cell of origin in various cancers [44,45]. Thus, the increased methylation in CIMP and T-ALL with respect to AML could stem from the higher levels of methylation in the lymphoid lineage, which are thought to be required to suppress key regulators of the myeloid lineage [26–28]. Nevertheless, data from differentiated cells revealed much inferior methylation levels, suggesting it might be a cancer-specific process. It cannot be ruled out that this methylation pattern occurs in intermediate progenitors not characterized in this study, such as early DN1 thymocytes, which exhibit higher methylation levels than prior or posterior stages [28]. Analysis of gene expression signatures also revealed expression of genes associated with T-cells and various progenitor subpopulations. On the other hand, open

chromatin, a reliable predictor of gene identity, indicates proximity to myeloid lineage. The inconsistency between different analyses is a likely consequence of phenotypic plasticity, but also of the heterogeneity of these leukemias, some of which appear as more myeloid (e.g. #DD166, #3491). The emerging conclusion from these results is that CIMPs are likely to stem from an early progenitor, possibly lymphoid-primed, but with the capability to differentiate into myelo-erythroid cell types. Of note, Zhang et al. reported that ETP-ALL is enriched for GMP and HSC gene sets, leading them to conclude it derives from stem cells, rather than ETPs as initially thought [22]. This is congruent with our observations in CIMP, once again underscoring the similarity between these entities, and suggests they both derive from very early lymphoid progenitors that precede the DN2 stage.

Aberrant methylation results in the silencing of several critical TFs involved in lineage specification, including *SPI1* [46], *CEBPA* [47], *IRF4* [48] and *IRF8* [48]. Interestingly, IRF4 and a few genes like *MAFB* (another inducer of monocytic maturation [49]) or *KLF4* are completely repressed in CIMP, whereas they remain active in some T-ALL cases. Some TFs involved in lymphopoiesis, such as *LEF1*, a nuclear mediator of WNT signaling that regulates early stages of thymocyte maturation [31] and repress CD4+ T-cell programs in CD8+ T-cells [50], are also silenced in CIMP leukemias. Deletion of *LEF1* results in the upregulation of non-T-lymphoid genes via genome reorganization [51], which could contribute to the mixed phenotype observed here. On the other hand, while SPI1 (PU.1) levels are lower than in AML, motif activity analysis reports significantly higher activity of SPI1 than in T-ALL (Figure S10D). Taken together, this underscores the notion that CIMP leukemias are an intermediate entity, in which hypermethylation of multiple TFs averts multiple lineage trajectories.

The loss of *CEBPA* appears to play an outsized role in orchestrating the transcriptional changes that lead to leukemogenesis, according to several lines of evidence. Firstly, unsupervised analyses of epigenomics data detected strong similarity between CIMP and AML subtypes in which *CEBPA* is either repressed or dysfunctional, namely t(8;21) AML and *CEBPA* DM AML. Secondly, motif activity analysis revealed dramatic changes in chromatin accessibility between CIMP and AML at regions containing C/EBP motifs, as well as between AML and T-ALL. Thirdly, *CEBPA* actively promotes myeloid differentiation at the expense of lymphoid commitment [52] by directly repressing the expression of T-cell genes [53]. *CEBPB* [54] and *CEBPD* [55] can rescue granulocytic defects in the absence of *CEBPA*, but CIMPs also exhibit reduced expression of both genes. In contrast with other members of the family, *CEBPG* is upregulated in these leukemias, possibly due to loss of repression by *CEBPA* [56], which is compatible with its role as a dominant inhibitor of other C/EBP proteins via heterodimerization [57]. Deletion of *CEBPA* or its +37 kb enhancer results in accumulation of immature myeloid blasts, yet no progression to AML [58,59], which has been attributed to the need for initial differentiation into early progenitors where leukemia can develop [60]. This

could explain why CIMPs are not myeloid leukemias, even though they express myeloid markers. Interestingly, the +42 kb enhancer that drives *CEBPA* expression in myeloid cells [61] is active in both CIMP and AML, but absent in T-ALL (Figure S11C-E). It is thus tempting to speculate that transformation took place in a cell type that would normally express *CEBPA*, once again pointing to an early progenitor that is only biased towards the lymphoid lineage, but retaining substantial multilineage priming. All things considered, loss of *CEBPA* without compensation by other members of its family is a likely key driver of the phenotype of CIMP leukemias. More broadly, these data further emphasizes the role of *CEBPA* is a critical lineage-determining factor that shapes the epigenetic landscape in hematopoiesis.

Aside from gene promoters, hypermethylation was also pronounced at binding sites for CTCF, which was accompanied by widespread loss of CTCF binding, particularly at peaks containing motifs with CpGs. Increased occupancy of CTCF has been reported in AML, particularly in hypomethylated promoters of myeloid TFs, whose expression becomes increased [62]. This process is apparently reverted in CIMP by hypermethylation. Since many of these sites co-located with loop anchors and TAD boundaries where CTCF stabilizes cohesion-mediated interactions, we expected a major impact on 3D genome organization, but this was not the case. A possible explanation is that CTCF loss does not necessarily abolish TADs. While total depletion of CTCF does lead to a global loss of TADs [63], alteration of a single CTCF site may [64,65] or may not [66,67] be sufficient to perturb a TAD boundary. This is partially due to the fact that many TAD boundaries harbor clusters of redundant CTCF binding sites that confer them resilience to small changes [68,69], but also to the existence of alternative mechanisms that preserve TAD boundaries [67]. Depletion of CTCF must be near complete for a significant impact on TAD insulation [63], which explains why the limited loss due to methylation changes results in mostly modest changes. On the other hand, although CTCF is present at the vast majority of TAD boundaries, it is only found at a small fraction of enhancer-promoter loops [40], which are frequently occupied instead by YY1 [70,71]. On the other hand, the reduced number of differential interactions identified may be a consequence of the limited sample size and resolution of this Hi-C dataset. The latter could be addressed in the future with Micro-C, which offers substantially higher resolution [72].

Nonetheless, hypermethylation-driven CTCF loss modulates 3D organization at specific loci, in keeping with previous studies [17,18]. This phenomenon may be complemented by changes in TFs like *LEF1*, which also modulates chromatin interactions [51]. A striking example is the disruption of several loops and TAD insulation at the *KLF4* locus, which presumably abolishes the interaction between its promoter and putative enhancers, inactive in CIMP. The lost CTCF binding site that normally stabilizes these loops is at the *KLF4* promoter, which is hypermethylated. Among its multiple roles in hematopoiesis [73], KLF4 is required for monocyte differentiation [74], whereas its downregulation is required for lineage commitment of T-cells [75]. During these processes, KLF4 stimulates the formation of open chromatin and directly establishes *de novo* chromatin loops independently of CTCF [76,77], possibly explaining changes

in the 3D structure of CIMP leukemias that do not co-occur with variations in CTCF binding. Inactivation of *KLF4* by promoter methylation is necessary for monocyte commitment [78] and has been previously reported in T-ALL [79] and chronic lymphocytic leukemia (CLL) [80]. Inhibition of T-cell genes by KLF4 impairs T-ALL progression [81]. Thus, the complete loss of *KLF4* in CIMPs potentially contributes to a blockade of the myeloid trajectory while enabling the expression of lymphoid genes. Notably, the expression of KLF4 in CLL can be rescued by inhibition of NOTCH1, which is frequently mutated in CLL [80]. As 43% of the CIMP cases also exhibit such activating mutations, targeting of NOTCH1 can be an attractive therapeutic avenue for these leukemias.

The mechanisms underlying aberrant methylation in CIMPs are uncertain. None of the recurrently mutated genes in this leukemia have any known involvement in the methylation machinery. However, expression of *TET2* was significantly downregulated relative to AML due to promoter hypermethylation, whereas *DNMT1*, *DNMT3A* and *DNMT3B* were slightly upregulated by either demethylation or gained chromatin interactions. That is, aberrant methylation could result from inactivation of demethylating enzymes coupled with an increase in *de novo* and maintenance methylation. As mentioned above, another likely possibility is that the methylation signature of CIMP leukemias is partially inherited from their cell of origin, explaining the similarity with a subset of T-ALLs. A distinctive feature of this aberrant methylation is that it preferentially localizes to “bivalent promoters”, in keeping with reports that bivalent promoters are susceptible to DNA hypermethylation in cancer [30]. One possible explanation is that H3K4me3, which protects bivalent promoters against DNA methylation by DNMT3A [82,83], is lost in these regions (Figure S11F). Moreover, DNMT3A has been reported to associate with PRC2, which could lead to hypermethylation of H3K27me3-marked domains in the absence of protective H3K4me3 [84]. This interaction could be facilitated by the lack of expression of DNMT3L (Table S9), which competes with DNMT3A and DNMT3B for interaction with PRC2 [85].

In conclusion, CIMP or *CEBPA*-silenced leukemias are a group of immature leukemias of ambiguous lineage very similar to ETP-ALLs. Their mixed phenotype and lineage infidelity are a reflection of a hybrid epigenomic landscape, with methylation patterns of lymphoid leukemias superimposed on an enhancer repertoire that preserves a large degree of myeloid potential (Figure S11G). The repression of *CEBPA* likely plays a key role in locking out the myeloid lineage, while the formation of new loops enables the expression of T-cell genes like *GATA3*. At the same time, silencing of other TFs required for T-cell commitment, such as KLF4, prevent terminal differentiation of T-cells. Further studies will be necessary to untangle the causal relationships between these multiple layers of epigenetic regulation. Taken together, this study provides a detailed picture of the unique epigenomic landscape of CIMP leukemias and identifies potential mechanisms driving their differentiation arrest. Furthermore, the data collected here constitute a useful epigenomic reference for subsequent studies in AML, T-ALL and leukemias with mixed phenotype.

ACKNOWLEDGEMENTS

The authors are indebted to the colleagues from the bone marrow transplantation group and the molecular diagnostics laboratory of the department of Hematology (E. Braakman, P. J. M. Valk) as well as collaborators from the department of Clinical Genetics (H.B. Beverloo) at the Erasmus University Medical Center for storage of samples, molecular and cytogenetic analysis of the leukemia cells. We are also grateful to Remco Hoogenboezem for assistance in the development of bioinformatics tools and to the rest of our colleagues in the department of Hematology for insightful input during work discussions. Furthermore, we thank Roberto Avellino for critically reading the manuscript. This work was funded by grants from the Dutch Cancer Foundation "Koningin Wilhelmina Fonds" and the Leukemia Lymphoma Society.

AUTHOR CONTRIBUTIONS

The study was designed and written by R.M.-L., C.G., B.W, R.D. and M.R. Wet lab experiments were performed by S. van H., A.S., L.S., D.G., D.H., S.P., N.D. and G.E. Bio-informatical analyses were conducted by R.M-L., C.G., J.V. and M.R. Patient samples and data were provided by C.G., A.S., A.R., R.A. W.H., C.T. and B.W.

My contributions to this work were: design of the study; processing and analysis of all high throughput sequencing data (WES, RNA-seq, ChIP-seq, ATAC-seq, Hi-C); data integration and visualization, including clustering analyses; data management and upload; interpretation of the results and writing of the manuscript.

CONFLICT OF INTEREST DISCLOSURE

The authors declare no competing interests.

METHODS

Patient material

Samples of AML, CIMP and T-ALL patients were collected from the biobanks of the Erasmus MC Hematology department (Rotterdam, The Netherlands) and the University Hospital Regensburg Internal Medicine department (Regensburg, Germany). Mononuclear cells were isolated from bone marrow or peripheral blood as described previously[7]. All patients provided written informed consent in accordance with the Declaration of Helsinki. Patient blasts were stored at -80°C in RLT+ buffer (Qiagen) and RNA and DNA was isolated using the AllPrep DNA/RNA mini kit (Qiagen, #80204) or stored in RNABee (Tel-Test, Inc.) and isolated by standard diagnostic procedures. RNA was converted into cDNA using the SuperScript II Reverse Transcriptase (Thermo Fischer Scientific) according to standard diagnostic procedures.

Statistics and data visualization

Statistical tests were conducted on R version 4.1.0 unless otherwise specified. Most plots were generated using the *ggplot2* R package, whereas heatmaps were created with *ComplexHeatmap* [86] and genomic regions were visualized with *plotgardener* [87].

Identification of functional regions

Putative enhancer and promoter regions were defined for genome-wide quantification of methylation. In both cases, they were defined by three complementary criteria: a) relative position to genes, b) telltale histone marks, and c) eRNA expression. For enhancer identification, we constructed a consensus collection of H3K27ac-marked regions present in 3 or more samples from the CIMP, AML, T-ALL and HSPC groups, excluding peaks that overlapped with 1kb windows around transcriptional start sites (TSS) by at least 5% of their width. This list was intersected with a collection of open chromatin regions derived from the same groups (detectable in at least 3 samples) and with putative enhancers detected by CAGE-seq by the FANTOM consortium[88] (*human_permissive_enhancers_phase_1_and_2.bed*). For promoter identification, we downloaded the CAGE peaks assigned to TSS by the FANTOM consortium (*hg19.cage_peak_phase1and2combined_tpm_ann.osc.txt.gz*), excluded those not expressed in healthy hematopoietic cells, and intersected them with H3K4me3 peaks from CD34+ cells obtained from ENCODE [89].

CpG islands, identified according to the original criteria of Gardiner-Garden and Frommer [90], were downloaded from the UCSC browser.

Quantification and differential analysis of peak-based data

Quantification of peak signal (MCIP-seq, ChIP-seq, ATAC-seq) was carried out with the *DiffBind* R package[91] as follows. First, all peaks were combined in a single master list using

the default settings of the package, keeping only peaks present in at least 2 samples and removing chromosomes not present in the primary assembly and unassigned sequences. Only peaks with a –log(q-value) higher than 10 as determined by MACS2 were considered. Overlapping peaks across multiple samples were combined into a single entry. The *dba.blacklist* function was used with the greylist argument set to false to remove only blacklisted regions. Then, reads mapping to this master list were counted for each sample, subtracting reads mapping to an input DNA samples processed in the same way.

For differential analysis, data were normalized using the trimmed mean of the M-values (TMM)[92] with the sum of reads in consensus peaks as the library size (argument normalize=DBA_NORM_TMM in *DiffBind*). These normalization factors and the raw counts were passed to DESeq2 (v1.34.0) with the *dba.analyze* command and differential regions were identified as those with a false discovery rate (FDR) < 0.5 by the Benjamini-Hochberg method [93]. Peaks were annotated to the closest gene with the *ChIPpeakAnno* package [94].

Clustering of transcriptional and epigenomics data

MCIP-seq, RNA-seq, ATAC-seq and CHIP-seq data were processed in the same way to identify relevant relationships between CIMP leukemias and other diseases. Briefly, raw counts were imported to DESeq2 with the *DESeqDataSetFromMatrix* function and transformed with *varianceStabilizingTransformation* in order to reduce the dependence of the variance from the mean. The 5000 regions or genes with the highest variance in transformed counts were selected for further analysis. Principal component analysis (PCA), multidimensional scaling (MDS), t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) were used for dimensionality reduction and visualized with *ggplot2*. Since results were very comparable across different strategies (Figure S1), only PCA and UMAP were used in the rest of the figures. Moreover, heatmaps of either Pearson correlation or Euclidean distances between samples were created with *ComplexHeatmap* [86].

Methyl-CpG immunoprecipitation sequencing (MCIp-seq) data generation and analysis

To measure methylation, we employed Methyl-CpG-immunoprecipitation (MCIP) a technique which relies on a fusion protein consisting of the methyl-binding domain (MBD) of MBD2 and the Fc portion of IgG1 to detect methylated regions, exploiting the natural preference of MBD for 5-methylcytosine (5-mC) [95]. MCIP-seq was performed using the EpiMark® Methylated DNA Enrichment Kit (NEB, Frankfurt, Germany) according to the manufacturer´s guidelines. In brief, genomic DNA was fragmented to an average size of 200 bp using the sonication system Covaris S220 (Covaris, Woburn, USA). Each sample (200ng) was incubated with 15µl MBD2-Fc/Protein A magnetic beads and incubated for 1h at room temperature. Unbound DNA was washed off with washing buffer containing 500mmol/L NaCl. Captured methylated DNA was recovered by adding 50µl DNAse free water and

incubation at 65°C for 15 minutes. The distribution of CpG methylation densities in both fractions (unmethylated and methylated) was controlled by qPCR using primers covering the imprinted SNRPN and a genomic region lacking CpGs (empty 6.2). Sequencing libraries were prepared with the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB) according to manufacturer's instructions. The quality of dsDNA libraries was analyzed using the High Sensitivity D1000 ScreenTape Kit (Agilent) and concentrations were determined with the Qubit dsDNA HS Kit (Thermo Fisher Scientific). Libraries were single-end sequenced on a HiSeq3000 (Illumina).

MCIP-seq reads were aligned to the human reference genome build hg19 with *bowtie* [96] (v1.1.1) and bigwig files were generated for visualization with *deepTools bamCoverage* [97] (v3.5.1). Peak calling was performed with MACS2 [98] (v2.1.2) using default settings and input DNA as a control. The resulting peaks were filtered against the ENCODE blacklisted regions [89]. Furthermore, a list of regions accessible by MCIP-seq was defined based on data from monocytes treated with the CpG Methyltransferase SssI. All peaks that did not overlap with this list of mappable regions were considered false positives and discarded using *bedtools intersect*[99]. Functional annotation of peaks was performed with the *ChIPseeker* (Figure 1A) and the *annotatr* [100] (Figure 1B) R packages.

RNA-seq data generation and differential expression analysis

Sample libraries were prepped using 500 ng of input RNA according to the KAPA RNA HyperPrep Kit with RiboErase (HMR) (Roche) using Unique Dual Index adapters (Integrated DNA Technologies, Inc.). Amplified sample libraries were paired-end sequenced (2x100 bp) on the Novaseq 6000 platform (Illumina) and aligned against the human genome (hg19) using STAR v2.5.4b [101].

Salmon [102] was used to quantify expression of individual transcripts, which were subsequently aggregated to estimate gene-level abundances with the R package *tximport* [103]. Human gene annotation derived from GENCODE [104] v30 was downloaded as a GTF file. Both gene- and transcript-level abundances were normalized to counts per million (CPM) for visualization in the figures of this paper. Differential gene expression analysis of count estimates from Salmon was performed with DESEq2 [93] v1.34.0.

Fusion gene detection

Fusion gene identification was carried out on RNA-seq reads by means of an ensemble of software tools, namely *STAR-Fusion* [105], *FusionCatcher* [106], *Arriba* [107], *Pizzly* [108], *JAFFA* [109] and *SQUID* [110]. Results from these tools were integrated with *fusion-reporter*, a python script developed for the *nf-core* framework of bioinformatics pipelines [111]. Fusion gene candidates previously found in studies of healthy tissues or involving partners in close proximity, as reported by the databases bundled with FusionCatcher, were discarded. Majority voting by

a minimum of 3 tools was employed to select the final fusion candidates per sample, which were then combined into a single master list.

The combined list of fusions was further annotated based on their presence in fusion gene databases (FusionGDB, COSMIC and Mitelman) or previous reports of that fusion in leukemia studies [112–115]. Fusions whose individual genes are involved in leukemia according to the Disgenet database [116] were also annotated. The master list and the leukemia-related annotations were visually represented with the *oncoPrint* function of the *ComplexHeatmap* R package.

Gene set enrichment analysis and identification of hematopoietic signatures

Gene set enrichment analysis (GSEA) [117] was computed with the *fgsea* R package using the multilevel splitting Monte Carlo approach to calculate p-values, with the settings minSize=15, maxSize=5000. We used the MSigDB C5 collection, containing GO terms, to investigate enrichment for gene functions and biological processes [118]. We also employed a customized MSigDB C2 collection, containing the version v7.5.1 of C2 plus several hematopoiesis-related datasets kindly provided by Dr. Charles Mullighan. Moreover, we added datasets derived from supervised comparisons between ETP-ALL and other T-ALLs [21,119], as well as a signature of leukemia induced in DN2 thymocytes mice by a retrovirus coexpressing *Myc* and *Bcl2* [120]. Both C2 and C5 were downloaded with the *msigdbr* R package.

To evaluate the potential cell of origin of CIMP leukemias, we analyzed the samples with single sample GSEA (ssGSEA) implemented as a part of the *GSVA* R package [121]. With that same goal, we employed *CIBERSORTx* [25], originally designed to dissect cell type proportions in a mixture on the basis of a signature matrix. Signature matrices were generated from single cell datasets obtained from the Human Cell Atlas [23] and the Atlas of Human Blood Cells [24].

ChIP-seq data generation and analysis

ChIP-seq data were generated with different antibodies targeted at histone marks (H3K27ac, H3K27me3) and CTCF. ChIP was performed as described previously with slight modifications [122]. Briefly, cells were crosslinked with 1% formaldehyde for 10 minutes at room temperature and the reaction was quenched with glycine at a final concentration of 0.125 M. Chromatin was sheared using the Covaris S220 focused-ultrasonicator to an average size of 250–350 bp. A total of 2.5 µg of antibody against H3K27ac (Abcam, ab4729) was added to sonicated chromatin of 2×10^6 cells and incubated overnight at 4 °C. Protein A sepharose beads (GE healthcare) were added to the ChIP reactions and incubated for 2 h at 4 °C. Beads were washed and chromatin was eluted. After crosslink reversal, RNase A and proteinase K treatment, DNA was extracted with the Monarch PCR & DNA Cleanup kit (NEB). Sequencing libraries were prepared with the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB) according to the manufacturer's instructions. The quality of dsDNA libraries was analyzed

using the High Sensitivity D1000 ScreenTape Kit (Agilent) and concentrations were assessed with the Qubit dsDNA HS Kit (Thermo Fisher Scientific). Libraries were single-end sequenced on a HiSeq3000 (Illumina).

ChIP-seq reads were aligned to the human reference genome build hg19 with *bowtie* [96] (v1.1.1) and bigwig files were generated for visualization with *deepTools bamCoverage* [123] (v3.5.1). For data with narrow read distributions (H3K27ac, CTCF), peak calling was performed with *MACS2*[98] (v 2.1.2) using default settings and the resulting peaks were filtered against the ENCODE blacklisted regions [89]. For H3K27me3, which is found in broad domains, peak calling was performed with EPIC2 [124].

ATAC-seq data and analysis

ATAC-seq was essentially carried out as described [125]. Briefly, prior to transposition the viability of the cells was assessed and 1×10^6 cells were treated in culture medium with DNase I (Sigma) at a final concentration of 200 U ml−1 for 30 minutes at 37 °C. After Dnase I treatment, cells were washed twice with ice-cold PBS, and cell viability and the corresponding cell count were assessed. 5 × 104 cells were aliquoted into a new tube and spun down at 500 × g for 5 minutes at 4 °C, before the supernatant was discarded completely. The cell pellet was resuspended in 50 µl of ATAC-RSB buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl2) containing 0.1% NP-40, 0.1% Tween-20, and 1% Digitonin (Promega), and was incubated on ice for 3 minutes to lyse the cells. Lysis was washed out with 1 ml of ATAC-RSB buffer containing 0.1% Tween-20. Nuclei were pelleted at 500 × g for 10 minutes at 4 °C. The supernatant was discarded carefully and the cell pellet was resuspended in 50 µl of transposition mixture (25 µl 2× tagment DNA buffer, 2.5 µl transposase (100 nM final; Illumina), 16.5 µl PBS, 0.5 µl 1% digitonin, 0.5 µl 10% Tween-20, 5 µl H2O) by pipetting up and down six times. The reaction was incubated at 37 °C for 30 minutes with mixing before the DNA was purified using the Monarch PCR & DNA Cleanup Kit (NEB) according to the manufacturer's instructions. Purified DNA was eluted in 20 µl elution buffer (EB) and 10 µl purified sample was objected to a ten-cycle PCR amplification using Nextera i7- and i5-index primers (Illumina). Purification and size selection of the amplified DNA were carried out with Agencourt AMPure XP beads. For purification the ratio of sample to beads was set to 1:1.8, whereas for size selection the ratio was set to 1:0.55. Purified samples were eluted in 15 µl of EB. Quality and concentration of the generated ATAC libraries were analyzed using the High Sensitivity D1000 ScreenTape Kit (Agilent) and libraries were sequenced paired-end on a NovaSeq (Illumina).

ATAC-seq reads were aligned to the human reference genome build hg19 with *bowtie2*[126] (v2.3.4.1), which is recommended for longer reads, and mitochondrial and duplicate reads were excluded. Bigwig files were generated as described above. Peak calling was also performed with MACS2[98] (v 2.1.2), but with the following settings*: --nomodel --shift 100 --extsize 200*. The resulting peaks were filtered against the ENCODE blacklisted regions [89].

Whole exome sequencing (WES)

The Genomic DNA Clean & Concentrator kit (ZYMO Research) was used to remove EDTA from the DNA samples. Sample libraries were prepared using 100 ng of input according to the KAPA HyperPlus Kit (Roche) using Unique Dual Index adapters (Integrated DNA Technologies, Inc.). Exomes were captured using the SeqCap EZ MedExome (Roche Nimblegen) according to SeqCap EZ HyperCap Library v1.0 Guide (Roche) with the xGen Universal blockers – TS Mix (Integrated DNA Technologies, Inc.). The amplified captured sample libraries were paired-end sequenced (2x100 bp) on the Novaseq 6000 platform (Illumina) and aligned to the hg19 reference genome using the Burrows-Wheeler Aligner (BWA)[127], v0.7.15-r1140.

Identification and analysis of small genetic variants

Single nucleotide variant (SNV) and small insertion/deletion (indel) detection was performed with a custom script that integrated variants called by multiple software tools, including HaplotypeCaller and *MuTecT2* from GATK v4.0.0 [128], *VarScan2* [129], *bcftools* [130], *Strelka2* [131] and *Pindel* [132]. A highly optimized in-house tool (*annotateBamStatistics*) was then used to compute the variant allele frequency (VAF) of every variant as well as position-specific metrics for such as strand bias, number of clipped reads or the number of alternative alignments (Table S2). The combined list of variants was subjected to stringent filtering to a remove low-quality positions, considering the following criteria:

a) strand bias between 0 and 1 for regions within the exome capture (+200 bp)
b) total sequencing depth of at least 8 reads and 4 for the variant allele
c) alignment quality 40 or more and base calling score 30 or more
d) fewer than 40% of reads mapping to a base other than the reference and alternative alleles
e) maximum of 10% of the reads with an alternative alignment or a superior alternative alignment score in *bwa* (XS)
f) removal of extremely long indels (500 bp or more)
g) removal of variants in simple repeats as detected by RepeatMasker [133](downloaded from UCSC)
h) removal of variants in highly repetitive genomic regions, as determined by 95% or more identity to another region in selfChain link files from UCSC
i) removal of clusters of at least 3 SNVs with a distance of less than 5 bp from each other

Furthermore, since we did not have control material for these patients, we selected mutations likely to be somatic among the variants identified by WES based on functional annotation by Annovar[134]. Thus, we first considered mutations complying with the following criteria: a) located in exons or in splicing acceptor regions, b) non-synonymous SNV or indels, c) with a VAF of at least 1%. Single nucleotide polymorphisms (SNPs) with a population frequency higher than 0.0002 were excluded unless they were reported in the

COSMIC database v94 [135] in at least 5 hematological cancers, or they were present in genes with frequent clonal hematopoiesis mutations (*DNMT3A, TET2, ASXL1*) [136]. Variants present in a healthy donor (though not a paired matched control) were also removed to further eliminate common variants and technical artifacts. Moreover, variants present in a blacklist of frequent non-somatic variants found in WES from AML and CD34+ cells were discarded. Finally, probable oncogenic variants were selected as those that fulfilled one or more of the following conditions: i) in COSMIC database; ii) frameshift, stopgain or startloss; iii) majority of damaging functional predictions by tools such as PolyPhen, SIFT, LRT and others.

Given the difficult interpretation of some of these variants, the resulting list was further reduced by selecting only genes previously reported in leukemia (Disgenet database [116]), cancer (COSMIC [135]) or relevant in hematopoiesis (GO term GO:0030097). This file (Table S2) was used as an input for the *oncoPrint* function of the *ComplexHeatmap* R package to show the distribution of mutations in this cohort.

Copy number alteration (CNA) detection

Copy number analysis on WES data was performed with *CNVkit* [137] v0.9.9 in two steps. First, a pooled reference was generated based on 12 datasets from healthy CD34+ cells (9 from adult bone marrow and 3 from cord blood). As suggested by the instructions of the program, 5 kb regions of poor mappability were excluded from the analysis. Subsequently, the reference was employed to compute log2 copy ratios and infer discrete copy number segments using the default settings of *CNVkit*. Finally, we derived absolute integer copy numbers of these segments with the function "cnvkit call" and copy number alterations (CNAs) were computed at the gene level with *cnvkit genemetrics*. Copy number data were summarized across all AML samples and represented as a heatmap with *ComplexHeatmap*. Scatter plots of specific regions such as *NF1* were created with *cnvkit scatter*.

These results were validated by orthogonal analyses with *CNV Radar*[138] on WES data and *Control-FREEC* [139] on input DNA sequencing data generated for the ChIP-seq and MCIp-seq experiments. For *CNV Radar*, common SNPs (db SNP 151) were annotated in the variants called by *bcftools call* with the *SnpSift* [140] tool, as prescribed by the instruction manual. This step ensures that the B-allele frequency (BAF) is only calculated with polymorphisms that are expected to be heterozygous, avoiding distortions introduced by potentially subclonal somatic mutations. A panel of non-matched normals was used as a control analogously to the previous analysis with *CNVkit*. *Control-FREEC* was run without controls in windows of 100,000 bp were used to compensate for the low sequencing depth of the files.

Hi-C data and analysis

Low-C was performed using 12k flow sorted cells as previously reported [141]. The following procedural modifications were made: 500U of Hind III-HF (NEB R3104) was used as the restriction enzyme instead of 100U of Mbol. The mock PCR amplification was monitored by qPCR instead of by using an agarose gel. This was done by removing the magnetic beads

from, and adding 20x sybr (Biotium 3100) to, a small aliquot of the PCR reaction after two cycles.

Hi-C data were first processed with *HiCUP* [142] v0.8.2, a pipeline for mapping and processing Hi-C data that removes technical artifacts and other invalid or uninformative di-tags. As part of this pipeline, the reads were aligned to the human reference genome build hg19 using *Bowtie2* [126] v2.3.4.1. Filtered di-tags were then extracted with the script *hicup2juicer* and subsequently binned with *juicer tools pre* [143] v1.22.01 at the default resolutions 2.5 Mb, 1 Mb, 500 Kb, 250 Kb, 100 Kb, 50 Kb, 25 Kb, 10 Kb, and 5 Kb. The resulting .hic files were used for visualization. Identification of TADs and loops was conducted for each group of leukemias with the *findTADsAndLoops.pl find* script of the HOMER suite with the parameters –res 5000 and –window 10000. Loops and TADs were then aggregated with the *merge2Dbed.pl* script and individual scores were calculated per each sample with *findTADsAndLoops.pl score* using the same settings as above. These scores were imported to *DESeq2* [93] (v1.24.0) with the *DESeqDataSetFromMatrix* function and transformed with *varianceStabilizingTransformation* for unsupervised analysis as described above for other epigenomics data. Differences in TADs and loop scores between conditions were computed with a Wald test using the *DESEq* function. The results were visualized with *plotgardener* [87].

Because each individual dataset was sequenced at relatively low depth of coverage (average = 398 M paired end reads, 300 M valid pairs and 155 M unique pairs), identification of structural features was conducted in aggregate as described above. The 4537 TADs and 9443 loops detected across all datasets were comparable to previous Hi-C results, such as 5975 domains and 6058 loops in K562 or 9274 domains and 9449 loops in GM12878 [41].

Integration of Hi-C data with gene expression and CTCF binding

TAD boundaries were defined as 5000 bp regions (same as the resolution used for TAD calling) centered on their borders. CTCF binding sites overlapping with those regions were identified, but only a single peak with the smallest FDR was kept for each boundary, depending on which comparison was conducted. Similarly, MCIP-seq peaks overlapping with boundaries were selected based on their FDR for each comparison. Differential expression of genes within TADs was also incorporated. This information is summarized in Tables S18 and S19, which were used to identify variable TADs with a) significant changes in CTCF binding in their boundaries, b) differentially expressed genes.

Loop anchors were defined according the coordinates provided by *HOMER*, with an extra padding of 5000 bp on each side to account for the resolution used in loop calling. Enhancers and promoters (see **Identification of enhancer and promoter regions**) in the vicinity of loop anchors were identified at a distance of 25,000 bp or less. Thus, we could select enhancer-promoter loops as those with an anchor close to an enhancer and a promoter on each side (Tables S20 and S21). For loop anchors attached to a promoter, differential expression of the corresponding gene between the conditions of interests was also retrieved.

REFERENCES

1. Shih AH, Abdel-Wahab O, Patel JP, Levine RL. The role of mutations in epigenetic regulators in myeloid malignancies. *Nat. Rev. Cancer*. 2012;12(9):599–612.
2. Liu Y, Easton J, Shao Y, et al. The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. *Nat. Genet.* 2017;49(8):1211–1218.
3. Gröschel S, Sanders MA, Hoogenboezem R, et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in Leukemia. *Cell*. 2014;157(2):369–381.
4. Ottema S, Mulet-Lazaro R, Beverloo HB, et al. Atypical 3q26/MECOM rearrangements genocopy inv(3)/t(3;3) in acute myeloid leukemia. *Blood*. 2020;136(2):224–234.
5. Mansour MR, Abraham BJ, Anders L, et al. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science (80-.).* 2014;346(6215):1373–1377.
6. Jost E, Lin Q, Weidner CI, et al. Epimutations mimic genomic mutations of DNMT3A in acute myeloid leukemia. *Leukemia*. 2014;28(6):1227–1234.
7. Valk PJM, Verhaak RGW, Beijen MA, et al. Prognostically Useful Gene-Expression Profiles in Acute Myeloid Leukemia. *N. Engl. J. Med.* 2004;350(16):1617–1628.
8. Wouters BJ, Jordà MA, Keeshan K, et al. Distinct gene expression profiles of acute myeloid/T-lymphoid leukemia with silenced CEBPA and mutations in NOTCH1. *Blood*. 2007;110(10):3706–3714.
9. Figueroa ME, Wouters BJ, Skrabanek L, et al. Genome-wide epigenetic analysis delineates a biologically distinct immature acute leukemia with myeloid/T-lymphoid features. *Blood*. 2009;113(12):2795–2804.
10. Gebhard C, Glatz D, Schwarzfischer L, et al. Profiling of aberrant DNA methylation in acute myeloid leukemia reveals subclasses of CG-rich regions with epigenetic or genetic association. *Leukemia*. 2019;33(1):26–36.
11. Kelly AD, Kroeger H, Yamazaki J, et al. A CpG island methylator phenotype in acute myeloid leukemia independent of IDH mutations and associated with a favorable outcome. *Leukemia*. 2017;31(10):2011–2019.
12. Jones PA, Baylin SB. The Epigenomics of Cancer. *Cell*. 2007;128(4):683–692.
13. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* 2019;20(10):590–607.
14. Jones PA. Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* 2012;13(7):484–492.
15. Yin Y, Morgunova E, Jolma A, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science (80-.).* 2017;356(6337):.
16. Ong C-T, Corces VG. CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.* 2014;15(4):234–246.
17. Wiehle L, Thorn GJ, Raddatz G, et al. DNA (de)methylation in embryonic stem cells controls CTCF-dependent chromatin boundaries. *Genome Res.* 2019;29(5):750–761.
18. Flavahan WA, Drier Y, Liau BB, et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*. 2016;529(7584):110–114.
19. Khan M, Siddiqi R, Naqvi K. An update on classification, genetics, and clinical approach to mixed phenotype acute leukemia (MPAL). *Ann. Hematol.* 2018;97(6):945–953.
20. Khoury JD, Solary E, Abla O, et al. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Myeloid and Histiocytic/Dendritic Neoplasms. *Leukemia*. 2022;36(7):1703–1719.
21. Coustan-Smith E, Mullighan CG, Onciu M, et al. Early T-cell precursor leukaemia: a subtype of very high-risk acute lymphoblastic leukaemia. *Lancet Oncol.* 2009;10(2):147–156.

22. Alexander TB, Gu Z, Iacobucci I, et al. The genetic basis and cell of origin of mixed phenotype acute leukaemia. *Nature*. 2018;562(7727):373–406.
23. Hay SB, Ferchen K, Chetal K, Grimes HL, Salomonis N. The Human Cell Atlas bone marrow single-cell interactive web portal. *Exp. Hematol.* 2018;68:51–61.
24. Xie X, Liu M, Zhang Y, et al. Single-cell transcriptomic landscape of human blood cells. *Natl. Sci. Rev.* 2021;8(3):.
25. Newman AM, Steen CB, Liu CL, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* 2019;37(7):773–782.
26. Bock C, Beerman I, Lien WH, et al. DNA Methylation Dynamics during In Vivo Differentiation of Blood and Skin Stem Cells. *Mol. Cell*. 2012;47(4):633–647.
27. Farlik M, Halbritter F, Müller F, et al. DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. *Cell Stem Cell*. 2016;19(6):808–822.
28. Ji H, Ehrlich LIR, Seita J, et al. Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature*. 2010;467(7313):338–342.
29. Bernstein BE, Mikkelsen TS, Xie X, et al. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell*. 2006;125(2):315–326.
30. Ohm JE, McGarvey KM, Yu X, et al. A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat. Genet.* 2007;39(2):237–242.
31. Okamura RM, Sigvardsson M, Galceran J, et al. Redundant regulation of T cell differentiation and TCRalpha gene expression by the transcription factors LEF-1 and TCF-1. *Immunity*. 1998;8(1):11–20.
32. Skokowa J, Cario G, Uenalan M, et al. LEF-1 is crucial for neutrophil granulocytopoiesis and its expression is severely reduced in congenital neutropenia. *Nat. Med.* 2006;12(10):1191–1197.
33. Wouters BJ, Löwenberg B, Erpelinck-Verschueren CAJ, et al. Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome. *Blood*. 2009;113(13):3088–3091.
34. Fasan A, Haferlach C, Alpermann T, et al. The role of different genetic subtypes of CEBPA mutated AML. *Leukemia*. 2014;28(4):794–803.
35. Pabst T, Mueller BU, Harakawa N, et al. AML1-ETO downregulates the granulocytic differentiation factor C/EBPα in t(8;21) myeloid leukemia. *Nat. Med.* 2001;7(4):444–451.
36. Felsenfeld G, Bell AC. Methylation of a CTCF-dependent boundary controls imprinted expressionof the *Igf2* gene. *Nature*. 2000;405(6785):482–485.
37. Hark AT, Schoenherr CJ, Katz DJ, et al. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature*. 2000;405(6785):486–489.
38. Wang H, Maurano MT, Qu H, et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.* 2012;22(9):1680–8.
39. Maurano MT, Wang H, Kutyavin T, Stamatoyannopoulos JA. Widespread site-dependent buffering of human regulatory polymorphism. *PLoS Genet.* 2012;8(3):e1002599.
40. Phillips-Cremins JE, Sauria MEG, Sanyal A, et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*. 2013;153(6):1281–1295.
41. Rao SSP, Huntley MH, Durand NC, et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. 2014;
42. Scamurra DO, Davey FR, Nelson DA, Kurec AS, Goldberg J. Acute leukemia presenting with myeloid and lymphoid cell markers. *Ann. Clin. Lab. Sci.* 1983;13(6):496–502.
43. Kim M, Costello J. DNA methylation: An epigenetic mark of cellular memory. *Exp. Mol. Med.* 2017;49(4):e322–e322.

44. Zhu T, Liu J, Beck S, et al. A pan-tissue DNA methylation atlas enables in silico decomposition of human tissue methylomes at cell-type resolution. *Nat. Methods*. 2022;19(3):296–306.
45. Bormann F, Rodríguez-Paredes M, Lasitschka F, et al. Cell-of-Origin DNA Methylation Signatures Are Maintained during Colorectal Carcinogenesis. *Cell Rep.* 2018;23(11):3407–3418.
46. Hohaus S, Petrovick M, Voso M, et al. PU.1 (Spi-1) and C/EBP alpha regulate expression of the granulocyte-macrophage colony-stimulating factor receptor alpha gene. *Mol. Cell. Biol.* 1995;15(10):5830–5845.
47. Avellino R, Delwel R. Expression and regulation of C/EBPα in normal myelopoiesis and in malignant transformation. *Blood*. 2017;129(15):2083–2091.
48. Tamura T, Nagamura-Inoue T, Shmeltzer Z, Kuwata T, Ozato K. ICSBP directs bipotential myeloid progenitor cells to differentiate into mature macrophages. *Immunity*. 2000;13(2):155–165.
49. Kelly LM, Englmeier U, Lafon I, Sieweke MH, Graf T. MafB is an inducer of monocytic differentiation. *EMBO J.* 2000;19(9):1987.
50. Xing S, Li F, Zeng Z, et al. Tcf1 and Lef1 transcription factors establish CD8+ T cell identity through intrinsic HDAC activity. *Nat. Immunol.* 2016;17(6):695–703.
51. Shan Q, Li X, Chen X, et al. Tcf1 and Lef1 provide constant supervision to mature CD8+ T cell identity and function by organizing genomic architecture. *Nat. Commun.* 2021;12(1):1–20.
52. Hasemann MS, Lauridsen FKB, Waage J, et al. C/EBPα Is Required for Long-Term Self-Renewal and Lineage Priming of Hematopoietic Stem Cells and for the Maintenance of Epigenetic Configurations in Multipotent Progenitors. *PLoS Genet.* 2014;10(1):.
53. Taskesen E, Avellino R, AlberichJorda M, et al. CEBPα Is a Transcriptional Repressor of T-Cell Related Genes Explaining the Myeloid/T-Lymphoid Features of CEBPα-Silenced AML. *Blood*. 2011;118(21):554.
54. Jones LC, Lin ML, Chen SS, et al. Expression of C/EBPbeta from the C/ebpalpha gene locus is sufficient for normal hematopoiesis in vivo. *Blood*. 2002;99(6):2032–2036.
55. Zhang L, Li J, Xu H, et al. Myc-Miz1 signaling promotes self-renewal of leukemia stem cells by repressing Cebpα and Cebpδ. *Blood*. 2020;135(14):1133–1145.
56. Alberich-Jordà M, Wouters B, Balastik M, et al. C/EBPγ deregulation results in differentiation arrest in acute myeloid leukemia. *J. Clin. Invest.* 2012;122(12):4490–504.
57. Parkin SE, Baer M, Copeland TD, Schwartz RC, Johnson PF. Regulation of CCAAT/enhancer-binding protein (C/EBP) activator proteins by heterodimerization with C/EBPgamma (Ig/EBP). *J. Biol. Chem.* 2002;277(26):23563–23572.
58. Zhang P, Iwasaki-Arai J, Iwasaki H, et al. Enhancement of hematopoietic stem cell repopulating capacity and self-renewal in the absence of the transcription factor C/EBPα. *Immunity*. 2004;21(6):853–863.
59. Avellino R, Mulet-Lazaro R, Havermans M, et al. Induced cell-autonomous neutropenia systemically perturbs hematopoiesis in Cebpa enhancer-null mice. *Blood Adv.* 2022;6(5):1406–1419.
60. Ye M, Zhang H, Yang H, et al. Hematopoietic Differentiation Is Required for Initiation of Acute Myeloid Leukemia. *Cell Stem Cell*. 2015;17(5):611–623.
61. Avellino R, Havermans M, Erpelinck C, et al. An autonomous CEBPA enhancer specific for myeloid-lineage priming and neutrophilic differentiation. *Blood*. 2016;127(24):2991–3003.
62. Mujahed H, Miliara S, Neddermeyer A, et al. AML displays increased CTCF occupancy associated with aberrant gene expression and transcription factor binding. *Blood*. 2020;136(3):339–352.
63. Nora EP, Goloborodko A, Valton AL, et al. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell*. 2017;169(5):930-944.e22.
64. Lupiáñez DG, Kraft K, Heinrich V, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*. 2015;161(5):1012–1025.
65. Guo Y, Xu Q, Canzio D, et al. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell*. 2015;162(4):900–910.

66. Rodríguez-Carballo E, Lopez-Delisle L, Zhan Y, et al. The HoxD cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes. *Genes Dev.* 2017;31(22):2264–2281.

67. Barutcu AR, Maass PG, Lewandowski JP, Weiner CL, Rinn JL. A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. *Nat. Commun.* 2018;9(1):1444.

68. Kentepozidou E, Aitken SJ, Feig C, et al. Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biol.* 2020;21(1):1–19.

69. Chang L-H, Ghosh S, Papale A, et al. A complex CTCF binding code defines TAD boundary structure and function. *bioRxiv*. 2021;2021.04.15.440007.

70. Weintraub AS, Li CH, Zamudio A V, et al. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell*. 2017;171(7):1573-1588.e28.

71. Beagan JA, Duong MT, Titus KR, et al. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res.* 2017;27(7):1139–1152.

72. Hsieh THS, Weiner A, Lajoie B, et al. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell*. 2015;162(1):108–119.

73. Ghaleb AM, Yang VW. Krüppel-like factor 4 (KLF4): What we currently know. *Gene*. 2017;611:27–37.

74. Feinberg MW, Wara AK, Cao Z, et al. The Kruppel-like factor KLF4 is a critical regulator of monocyte differentiation. *EMBO J.* 2007;26(18):4138–4148.

75. Liu X, Wen X, Liu H, Xiao G. Downregulation of the transcription factor KLF4 is required for the lineage commitment of T cells. *Cell Res.* 2011;21(12):1701–1710.

76. Di Giammartino DC, Kloetgen A, Polyzos A, et al. KLF4 is involved in the organization and regulation of pluripotency-associated three-dimensional enhancer networks. *Nat. Cell Biol.* 2019;21(10):1179–1190.

77. Wei Z, Gao F, Kim S, et al. Klf4 Organizes Long-Range Chromosomal Interactions with the Oct4 Locus in Reprogramming and Pluripotency. *Cell Stem Cell*. 2013;13(1):36–47.

78. Karpurapu M, Ranjan R, Deng J, et al. Krüppel Like Factor 4 Promoter Undergoes Active Demethylation during Monocyte/Macrophage Differentiation. *PLoS One*. 2014;9(4):93362.

79. Shen Y, Park CS, Suppipat K, et al. Inactivation of KLF4 promotes T-cell acute lymphoblastic leukemia and activates the MAP2K7 pathway. *Leukemia*. 2017;31(6):1314–1324.

80. Filarsky K, Garding A, Becker N, et al. Krüppel-like factor 4 (KLF4) inactivation in chronic lymphocytic leukemia correlates with promoter DNA-methylation and can be reversed by inhibition of NOTCH signaling. *Haematologica*. 2016;101(6):e249–e253.

81. Li W, Jiang Z, Li T, et al. Genome-wide analyses identify KLF4 as an important negative regulator in T-cell acute lymphoblastic leukemia through directly inhibiting T-cell associated genes. *Mol. Cancer*. 2015;14(1):.

82. Otani J, Nankumo T, Arita K, et al. Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX-DNMT3-DNMT3L domain. *EMBO Rep.* 2009;10(11):1235–1241.

83. Ooi SKT, Qiu C, Bernstein E, et al. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature*. 2007;448(7154):714–717.

84. Viré E, Brenner C, Deplus R, et al. The Polycomb group protein EZH2 directly controls DNA methylation. *Nature*. 2006;439(7078):871–874.

85. Neri F, Krepelova A, Incarnato D, et al. Dnmt3L antagonizes DNA methylation at bivalent promoters and favors DNA methylation at gene bodies in ESCs. *Cell*. 2013;155(1):121.

86. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*. 2016;32(18):2847–2849.

87. Kramer NE, Davis ES, Wenger CD, et al. Plotgardener: Cultivating precise multi-panel figures in R. *Bioinformatics*. 2022;38(7):2042–2045.

88. Andersson R, Gebhard C, Miguel-Escalada I, et al. An atlas of active enhancers across human cell types and tissues. *Nature*. 2014;507(7493):455–461.

89. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57–74.
90. Gardiner-Garden M, Frommer M. CpG Islands in vertebrate genomes. *J. Mol. Biol.* 1987;196(2):261–282.
91. Ross-Innes CS, Stark R, Teschendorff AE, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*. 2012;481(7381):389.
92. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 2010;11(3):R25.
93. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):550.
94. Zhu LJ, Gazin C, Lawson ND, et al. ChIPpeakAnno: A Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*. 2010;11(1):1–10.
95. Gebhard C, Schwarzfischer L, Pham TH, et al. Genome-wide profiling of CpG methylation identifies novel targets of aberrant hypermethylation in myeloid leukemia. *Cancer Res.* 2006;66(12):6118–6128.
96. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10(3):R25.
97. Ramírez F, Ryan DP, Grüning B, et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* 2016;44(W1):W160-5.
98. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 2008;9(9):R137.
99. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–842.
100. Cavalcante RG, Sartor MA. annotatr: genomic regions in context. *Bioinformatics*. 2017;33(15):2381–2383.
101. Dobin A, Davis CA, Schlesinger F, et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29(1):15–21.
102. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*. 2017;14(4):417–419.
103. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*. 2016;4:1521.
104. Frankish A, Diekhans M, Ferreira AM, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 2019;47(D1):D766–D773.
105. Haas BJ, Dobin A, Li B, et al. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.* 2019;20(1):1–16.
106. Nicorici D, ¸atalan MS, Edgren H, et al. FusionCatcher a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *bioRxiv*. 2014;011650.
107. Uhrig S, Ellermann J, Walther T, et al. Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res.* 2021;31(3):448–460.
108. Melsted P, Hateley S, Joseph IC, et al. Fusion detection and quantification by pseudoalignment. *bioRxiv*. 2017;166322.
109. Davidson NM, Majewski IJ, Oshlack A. JAFFA: High sensitivity transcriptome-focused fusion gene detection. *Genome Med.* 2015;7(1):1–12.
110. Ma C, Shao M, Kingsford C. SQUID: Transcriptomic structural variation detection from RNA-seq. *Genome Biol.* 2018;19(1):1–16.
111. Ewels PA, Peltzer A, Fillinger S, et al. The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* 2020;38(3):276–278.
112. Chen X, Wang F, Zhang Y, et al. Fusion gene map of acute leukemia revealed by transcriptome sequencing of a consecutive cohort of 1000 cases in a single center. *Blood Cancer J.* 2021;11(6):1–10.

113. Chen B, Jiang L, Zhong ML, et al. Identification of fusion genes and characterization of transcriptome features in T-cell acute lymphoblastic leukemia. *Proc. Natl. Acad. Sci. U. S. A.* 2017;115(2):373–378.

114. Wang Y, Wu N, Liu D, Jin Y. Recurrent Fusion Genes in Leukemia: An Attractive Target for Diagnosis and Treatment. *Curr. Genomics*. 2017;18(5):.

115. Huret JL, Ahmad M, Arsaban M, et al. Atlas of genetics and cytogenetics in oncology and haematology in 2013. *Nucleic Acids Res.* 2013;41(D1):.

116. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* 2020;48(D1):D845–D855.

117. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 2005;102(43):15545–15550.

118. Liberzon A, Subramanian A, Pinchback R, et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011;27(12):1739–1740.

119. Zhang J, Ding L, Holmfeldt L, et al. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature*. 2012;481(7380):157–163.

120. Riemke P, Czeh M, Fischer J, et al. Myeloid leukemia with transdifferentiation plasticity developing from T-cell progenitors. *EMBO J.* 2016;35(22):2399–2416.

121. Barbie DA, Tamayo P, Boehm JS, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*. 2009;462(7269):108.

122. Pham TH, Benner C, Lichtinger M, et al. Dynamic epigenetic enhancer signatures reveal key transcription factors associated with monocytic differentiation states. *Blood*. 2012;119(24):e161–e171.

123. Ramírez F, Ryan DP, Grüning B, et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* 2016;44(W1):W160–W165.

124. Stovner EB, Sætrom P. Epic2 efficiently finds diffuse domains in ChIP-seq data. *Bioinformatics*. 2019;35(21):4392–4393.

125. Corces MR, Trevino AE, Hamilton EG, et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods*. 2017;14(10):959–962.

126. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. 2012;9(4):357–359.

127. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760.

128. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297–1303.

129. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012;22(3):568–576.

130. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987–2993.

131. Saunders CT, Wong WSW, Swamy S, et al. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*. 2012;28(14):1811–1817.

132. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. 2009;25(21):2865–2871.

133. Tarailo-Graovac M, Chen N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinforma.* 2009;25(1):4.10.1-4.10.14.

134. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38(16):e164.

135. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* 2019;47(D1):D941–D947.

136. Jaiswal S, Ebert BL. Clonal hematopoiesis in human aging and disease. *Science (80-.).* 2019;366(6465):.

137. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput. Biol.* 2016;12(4):e1004873.

138. Soong D, Stratford J, Avet-Loiseau H, et al. CNV Radar: An improved method for somatic copy number alteration characterization in oncology. *BMC Bioinformatics*. 2020;21(1):.

139. Boeva V, Popova T, Bleakley K, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*. 2012;28(3):423.

140. Cingolani P, Patel VM, Coon M, et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front. Genet.* 2012;3(MAR):35.

141. Díaz N, Kruse K, Erdmann T, et al. Chromatin conformation analysis of primary patient tissue using a low input Hi-C method. *Nat. Commun. 2018 91*. 2018;9(1):1–13.

142. Wingett S, Ewels P, Furlan-Magaril M, et al. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research*. 2015;4:.

143. Durand NC, Shamim MS, Machol I, et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* 2016;3(1):95–98.

144. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood*. 2016;127(20):2391–2405.

145. Novershtern N, Subramanian A, Lawton LN, et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*. 2011;144(2):296–309.

146. Feingold EA, Good PJ, Guyer MS, et al. The ENCODE (ENCyclopedia of DNA Elements) Project. *Science (80-.).* 2004;306(5696):636–640.

147. Speir ML, Zweig AS, Rosenbloom KR, et al. The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.* 2016;44(D1):D717-25.

SUPPLEMENTARY INFORMATION

SUPPLEMENTARY RESULTS

The mutational landscape of CIMPs is similar to that of ETP-ALL

Although CIMP cases were originally diagnosed as AML, the mixed lymphoid and myeloid phenotype of these leukemias calls for a revised classification. To evaluate their relationship with other leukemias at the genetic level, we compiled data from published studies on AML [1,2], T-ALL [3–6], ETP-ALL [5–9] and T/M MPAL [10,11] (Table S6, Table S7). Genes like *NOTCH1*, *PHF6*, *JAK3* or *DNM2* were more often mutated in CIMP than in AML (p-value < 0.05, Fisher's exact test), but not than in T-ALL, whereas *DNMT3A*, *IDH2* and *RUNX1* exhibited mutation frequencies more similar to AML than to T-ALL. Mutation frequencies were generally comparable to T/M MPAL and especially ETP-ALL, although this assessment might be limited by the small sample size of the CIMP cohort. A significant difference was found for *MED12*, which was mutated in 36% of the CIMP cases and only 5% of all ETP-ALL and 1% of T/M MPAL. Interestingly, only two CIMP patients harbored lesions in *CHEK2*, a gene involved in DNA damage response. Although rare in *de novo* leukemia, germline *CHEK2* mutations have been reported to increase predisposition to various hematological malignancies, often following treatment for solid cancer [12–15]. Data on gene fusion were less conclusive, as there were no recurrent fusions identified in CIMP cases and studies in other cohorts yielded contradictory results (Table S8).

The frequency of CNAs in CIMPs (average, 13/genome) was higher than in AML (average, 6.3/genome), but lower than in T-ALL (average, 24.1/genome). However, aneuploidies were found much more frequently in the CIMP group (7/14) than in both AML (5/44) and T-ALL (5/14). This fraction was also much larger than the 20% of AML with aneuplodies reported in the literature [16], and frequent CNAs observed in AML, such as gain of chr8 or loss of chr7, were not particularly overrepresented in this cohort [17,18]. Altogether, this suggests unexpected genetic instability in these leukemias, which may be linked to alterations, genetic or epigenetic, in genes involved in DNA repair.

Even though no single gene was mutated in all CIMP cases, some mutations were present in more than 30% of the patients, including *NOTCH1*, *PHF6*, *NF1, MED12* and *WT1*. While some of those are found at similar frequencies in ETP-ALL, lesions in *NF1* and *MED12* were uniquely abundant in CIMP leukemias. *MED12* is a subunit of the Mediator complex, which plays critical functions in the regulation of transcription at multiple levels, including initiation, pausing and elongation [105]. The Mediator complex has been implicated in short-range interactions in collaboration with cohesin [42,106], but recent studies indicate it may act as a functional rather than an architectural bridge between enhancers and promoters [107,108]. Importantly, MED12 is essential for HSC function by cooperating with p300 in the

maintenance of enhancer activity [109]. Together with other mutations affecting epigenetic modifiers, *MED12* lesions can thus contribute to the abnormal epigenetic state of these leukemias. The recurrence of NF1 copy number losses was also noteworthy. Inactivation of NF1 results in RAS overactivation and increased blast colony formation and has been associated with poor prognosis in AML [110,111], but is also detected in T-ALL with and without neurofibromatosis [112,113]. The specific contribution of these mutations to the unique epigenetic phenotype of these leukemias remains to be elucidated.

Analysis of active and inactive chromatin

To further understand the epigenetic makeup of CIMP leukemias, we conducted additional analyses of the ATAC-seq and ChIP-seq data mentioned above. Contrary to methylation and expression data, there were no large differences between CIMP and either AML or T-ALL in terms of both open chromatin (Figure S9A) and H3K27ac deposition (Figure S9D). Unexpectedly, however, chromatin was slightly more open at promoters in CIMPs relative to AML, despite the widespread hypermethylation leading to gene silencing. Supervised comparisons similarly revealed that even numbers of promoters and enhancers are either active or inactive in CIMP with respect to AML or T-ALL (Figures S9B, S9E).

To draw more meaningful conclusions, we conducted GSEA on genes located in the vicinity of these variable peaks, ranked according to their differential binding in each comparison. CIMPs exhibited marked depletion of open chromatin at the targets of the PRC complex, i.e. regions marked by H3K27me3, relative to both AML and CD34+ cells (Figure S9C). This is in keeping with the preferential DNA methylation at these same regions, which become increasingly closed as well. Remarkably, CIMPs displayed open chromatin at genes normally expressed in AML when compared to T-ALL, whereas genes involved in T-cell differentiation were closed. Similar observations were derived from GSEA on H3K27ac (Figure S9F), but these same T-cell gene sets were more active than in CD34+ cells. Likewise, H3K27ac GSEA in CIMP vs AML revealed upregulation of T-cell sets and downregulation of AML ones. This is consistent with the epigenetic ambiguity of these leukemias and the notion that their differentiation is blocked at an intermediate stage, thus preventing a full commitment to either the T-lymphoid or the myeloid lineages.

Finally, we also also generated H3K27me3 ChIP-seq for a few CIMP and AML samples to investigate whether the increases in DNA methylation are accompanied by increased deposition of H3K27me3. Interestingly, there was a slight decrease in global H3K27me3 levels (Figure S9G), confirmed by supervised comparisons carried out with DiffBind (Figure S9H). Therefore, the establishment of DNA methylation does not require spreading of H3K27me3 to additional regions. GSEA revealed enrichment of H3K27me3 at genes involved in mitosis and NOTCH signalling, and depletion at genes downregulated in various types of AML (*NPM1*-mutant, RUNX1-ETO fusions, etc.). However, none of these results were significant with a FDR < 0.05, possibly due to the small size of the dataset.

Analysis of motif activity in chromatin accessibility data

The regulatory networks driving a differentiation block in CIMP were investigated with *chromVAR*, a tool that estimates TF motif activity by computing bias-corrected deviations in chromatin accessibility at motif-containing peaks relative to the expectation. Sample clustering based on the Pearson correlation between their motif accessibility scores grouped most CIMP with T-ALL (Figure S10A), in contrast with the analysis of global accessibility data, which segregated T-ALL from CIMPs (Figure S2H). In other words, while open chromatin regions preferentially contain binding sites for the same TFs that are active in T-ALL, the distribution of all accessible regions is not exclusively myeloid. Indeed, comparisons between CIMP and T-ALL revealed a large number of differential peaks genome-wide (Figure S9B). This is in line with the hypothesis that CIMPs derive from a lymphoid-biased progenitor with multilineage capacity that have acquired an incomplete myeloid differentiation program.

The most variable motifs across the entire leukemia cohort belonged to the JUN and FOS families of proto-oncogenes, which are positive regulators of myeloid differentiation [19] and are overexpressed in certain subtypes of AML [20,21]. Besides, several members of the C/EBP and GATA families were also very variable, consistently with the important roles in differentiation of genes like *GATA2* [22] or *CEBPA* [23]. Some of the most variable motifs, as well as other TFs of interest, were evaluated for synergy (Figure S9E). CTCF was highly antagonistic with KLF4, CEBPA, SPI1 and FOS, but positively associated with LEF1. Notably, both KLF4 [24,25] and LEF1 [26] are involved in structural reorganization of the genome, like CTCF, with KLF4 forming loops independently of CTCF. CEBPA and FOS were also strongly synergistic, in line with observations that JUN forms heterodimers with CEBPA that direct monocytic differentiation more potently than either of the two TFs alone [27]. However, high JUN expression has also been reported to inhibit CEBPA binding in AML, indicating a possible competitive behaviour as well [28].

Supervised comparisons of motif activity revealed large differences in C/EBP between CIMP and AML, as well as between AML and T-ALL (Figure S10D). This suggests C/EBP activity might be critical for the loss of myeloid potential in CIMP leukemias, in line with previous reports of the essential function of *CEBPA* in the establishment of the myeloid trajectory [23]. Interestingly, other TFs were significantly more active in CIMP than in T-ALL, including *SPI1* (PU.1), which induces myeloid commitment at high levels[29] and whose downregulation is necessary for terminal T-cell maturation[30], but also BACH1, which promotes B-cell development at the expense of the myeloid lineage [31].

Additional examples of altered 3D genome structure

In order to detect changes in 3D genome structure that lead to alterations in gene expression, we identified variable TADs and loops that contained or overlapped differentially expressed genes. Most of these differences were observed between CIMP and AML, in line with the notion that CIMP leukemias may derive from a lymphoid progenitor.

We first investigated whether any TFs silenced by promoter methylation exhibited concomitant changes in chromatin interactions. Only three of these TFs were found in loops with decreased intensity in the CIMP group: *PKNOX2*, *KLF4* (Figure 8G), *CEBPD* (Figure 8H). In the latter two, the loss of interaction was accompanied by reduced CTCF binding and gain of methylation at the *CEBPD* and *KLF4* promoters. On the other hand, 7 downregulated genes were found in ΔTADs, including again *CEBPD* and *KLF4,* but also other TFs like *TAL1*, *IRF8* (Figure S8J) and *MAFB*. In view of these observations, changes in 3D structure may contribute to silencing of TFs driving a differentiation block, but they are dispensable for a process that is largely driven by promoter silencing.

Next, we conducted an unbiased survey of ΔTADs (Tables S18-S19) and DIs (Tables S20-S21) with associated changes in CTCF binding and potential implications for gene expression. Aside from the examples described in the main text, the *IRF8* TAD exhibits reduced insulation accompanied by a loss of CTCF binding on the right boundary. Similarly, the TAD containing *GASK1B* was partially lost, potentially leading to downregulation of this gene due to the lack of interaction with a proximal enhancer (Figure S8K). On the other hand, the *ANGPT2* TAD became strongly insulated, possibly resulting in upregulation of this gene (Figure S8L). Gain of interaction between *GATA3* and a putative enhancer that is only active in CIMP and not in AML possibly lead to overexpression of *GATA3* (Figure 8I). Likewise, a loop involving the promoter of *DNMT3B* (Figure S8M) In contrast, the loss of interaction between *IL6* and putative upstream enhancer elements may explain the downregulation of this gene, whose promoter was unmethylated (Figure S8M).

There were only 3 ΔTADs and DIs between CIMP and T-ALL with reduced CTCF binding, one of which was loss of insulation at the TAD containing *ASCC1* (Figure S8O). This limited chromatin remodelling is in keeping with the notion that these leukemias originate from a lymphoid-biased cell. On the other hand, there were also fewer T-ALL replicates, which decreased the statistical power to detect such changes. The 8 differential enhancer-promoter loops between CIMP and T-ALL did not exhibit a clear correlation with gene expression (Figure 8F). Among those DIs was a loss of interaction between the promoter of the longer *DIPK1A* isoforms avnd downstream exonic regions of the same gene, which is downregulated in CIMPs (Figure S8P). This downregulation is accompanied by a loss of H3K27ac at said promoter. Another example is the gained interaction between *RCC2* and distal regulatory elements close to *SDHB*, which could possibly contribute to the overexpression of that gene (Figure S8Q).

SUPPLEMENTARY FIGURES

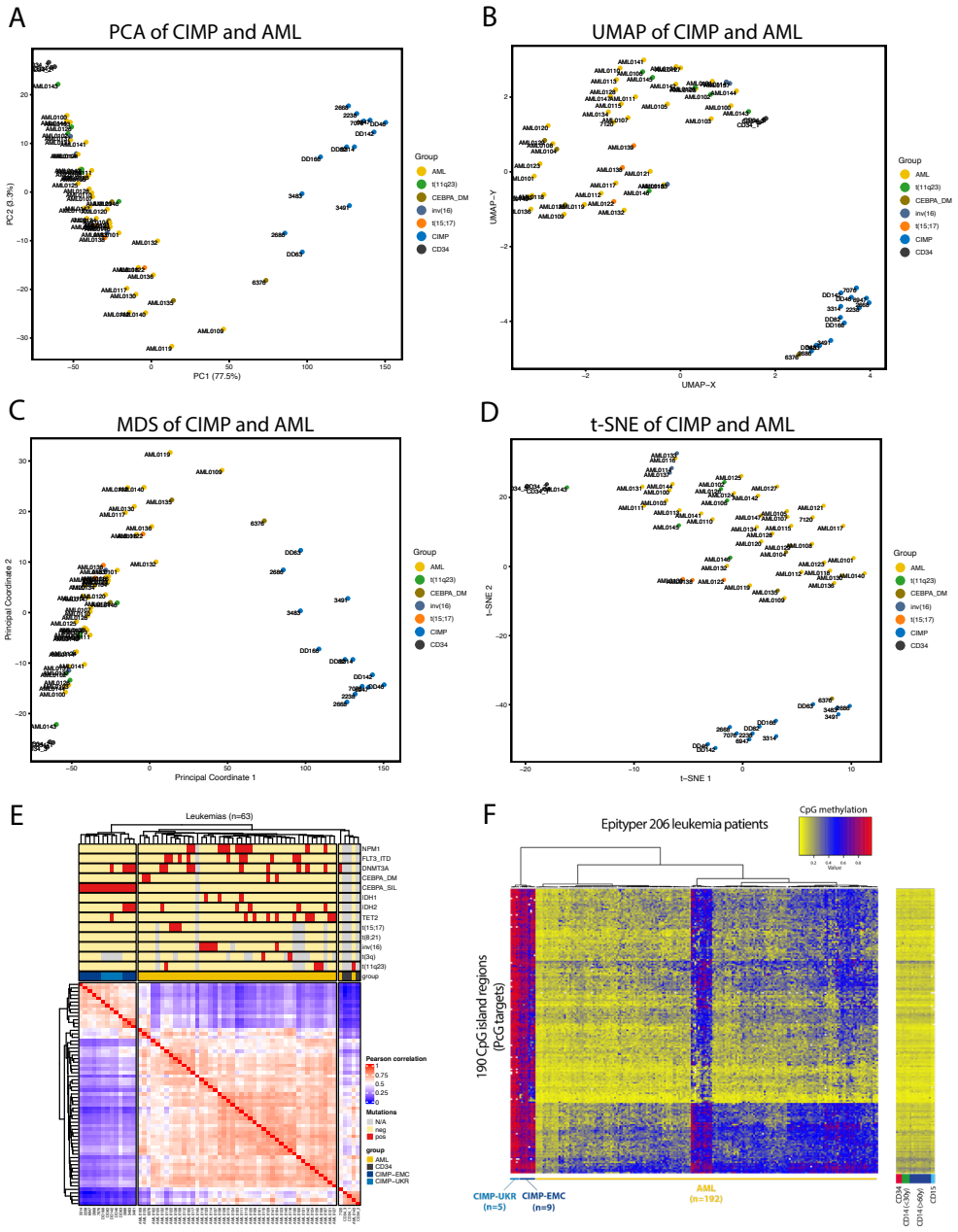


Figure S1. Dimensionality reduction of MCIP-seq data with different strategies. Methylation similarly separates CIMP from AML and CD34+ HSPCs using PCA (**A**), UMAP (**B**), MDS (**C**) or t-SNE (**D**). Relevant AML subgroups known to exhibit distinct patterns of gene expression are highlighted. **E.** Heatmap of the 3000 most variable MCIP-seq regions across all samples, displaying their Z-scores and clustered by Euclidean distance.

Figure S2

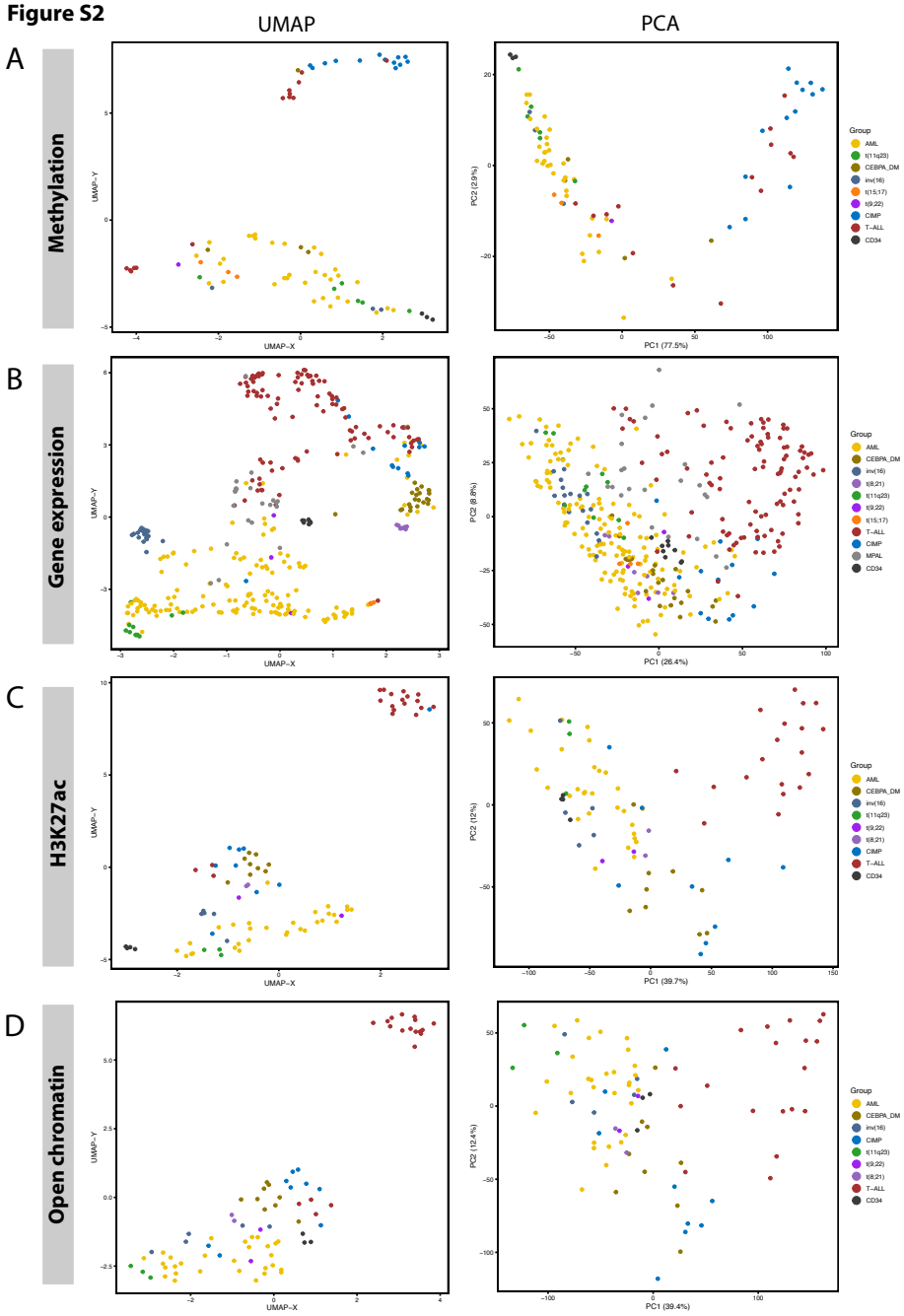


Figure S2. Epigenetic and transcriptional landscape of CIMP, AML, T-ALL and CD34+ cells. Dimensionality reduction with either UMAP or PCA of various types of epigenomics data (**A-D**) in AML, CIMP, T-ALL and CD34+ HSPCs. Relevant AML subgroups known to exhibit distinct patterns of gene expression are highlighted, revealing proximity between CIMP and *CEBPA* DM AML.

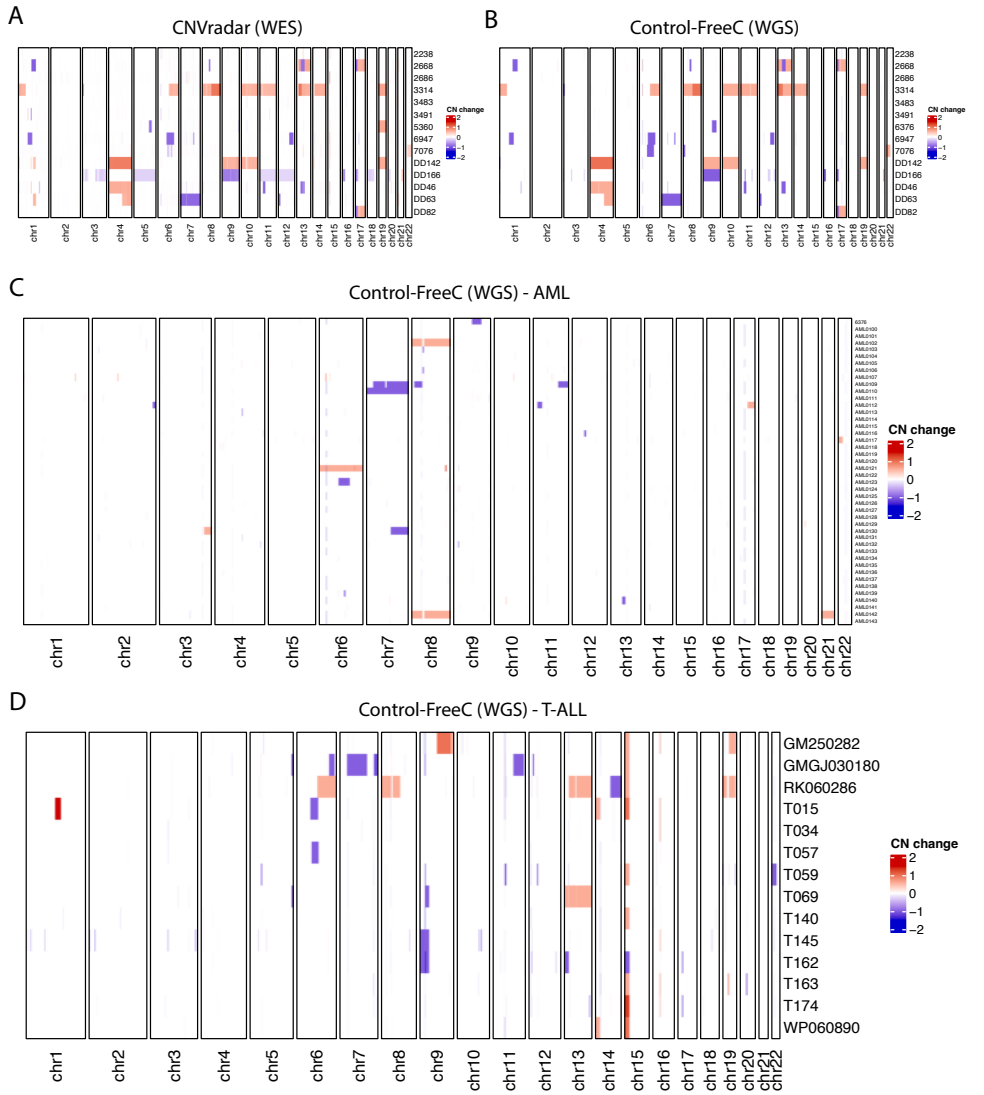


Figure S3. Validation of copy number analysis and comparison with other leukemias. Copy number alterations detected by different algorithms were represented as a heatmap where red indicates a copy number gain and blue a copy number loss. **A.** Reanalysis of WES data by CNV Radar. B. Validation of results from WES data in input DNA-sequencing with Control-FREEC. **C.** Copy number analysis of AML input data with Control-FREEC. **D.** Copy number analysis of AML input data with Control-FREEC.

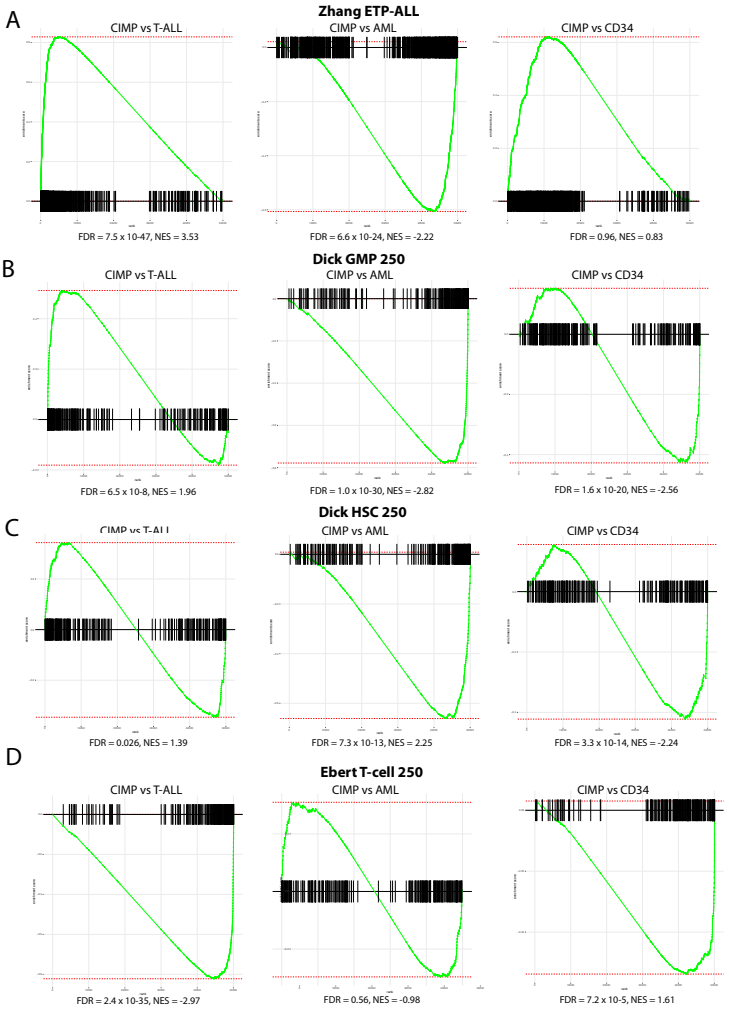


Figure S4. GSEA and signature analysis reveals that CIMPs have an early cell of origin. A-D. GSEA enrichment plots from comparisons between AML and T-ALL (left), AML (center) and CD34+ cells (right). The false discovery rate (FDR) and the normalized enrichment score (NES) are indicated underneath. **E.** Heatmap with the results of single sample GSEA (ssGSEA) for each individual sample in the cohort, considering only hematopoietic datasets derived from various hematopoietic fractions by the Ebert group [32]. The ssGSEA values were normalized as Z-scores, with blue corresponding to negative values (less enrichment) and red to positive values (more enrichment). No clustering was applied, samples are grouped by the leukemia types and subtypes they belong to, indicated at the top. **F.** Heatmap with the ssGSEA results from E, but averaged for each leukemia type (CEBPA DM AML is shown separately given its similarity to the CIMP group in previous analyses) and normalized as Z-scores. **G.** Same as E., but using datasets derived from a study by the Dick group [33]. **H.** Same as G., but using datasets from the Dick group. **I.** Heatmaps showing the cellular composition inferred by CIBERSORTx for each sample based on a signature derived from the Atlas of Human Blood Cells [34] (same dataset as in Figure 4C). The analysis was performed on a mixture matrix containing RNA-seq raw counts from multiple leukemia subgroups. The results are presented for each individual sample (columns) following Z-score normalization by cell type (rows). Hierarchical clustering was performed for both rows and columns using Manhattan distances. **J.** Same as I., but using a dataset derived from the Human Cell Atlas [35] for deconvolution. **K.** Box plot displaying the CIBERSORTx scores using a signature matrix derived from the Human Cell Atlas.

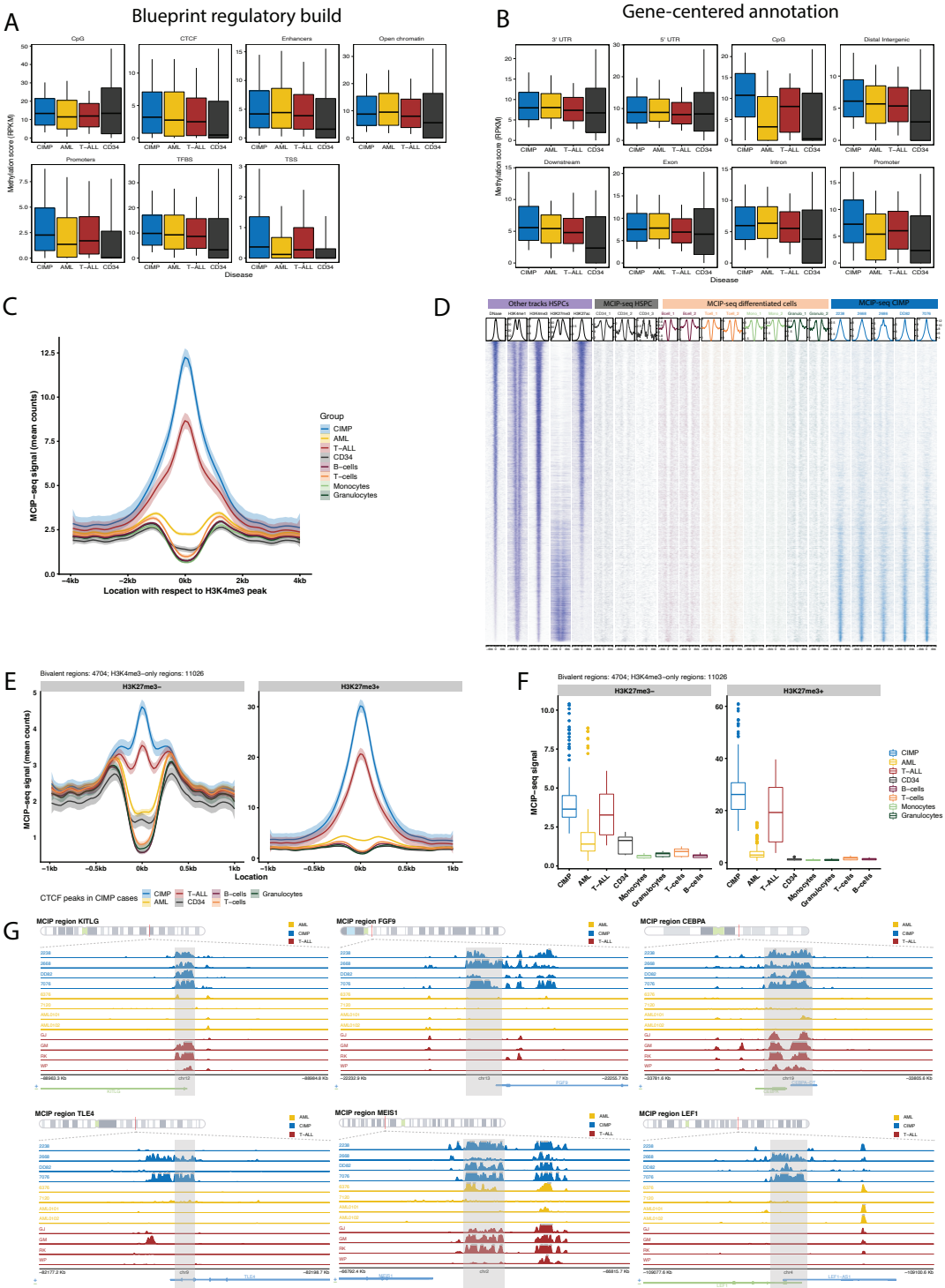

A
Blueprint regulatory build
CpG
CTCF
Enhancers
Open chromatin
Promoters
TFBS
TSS
Methylation score (RPKM)
CIMP AML T−ALL CD34
Disease
B
Gene-centered annotation
3' UTR
5' UTR
CpG
Distal Intergenic
Downstream
Exon
Intron
Promoter
Methylation score (RPKM)
Disease
C
MCIP−seq signal (mean counts)
Location with respect to H3K4me3 peak
−4kb −2kb 0kb 2kb 4kb
Group
CIMP
AML
T−ALL
CD34
B−cells
T−cells
Monocytes
Granulocytes
D
Other tracks HSPCs
MCIP-seq HSPC
MCIP-seq differentiated cells
MCIP-seq CIMP
E
Bivalent regions: 4704; H3K4me3−only regions: 11026
H3K27me3−
H3K27me3+
MCIP−seq signal (mean counts)
Location
−1kb −0.5kb 0kb 0.5kb 1kb
CTCF peaks in CIMP cases
CIMP AML T−ALL CD34 B−cells T−cells Granulocytes
F
Bivalent regions: 4704; H3K4me3−only regions: 11026
H3K27me3−
H3K27me3+
MCIP−seq signal
CIMP AML T−ALL CD34 Monocytes Granulocytes T−cells B−cells
G
MCIP region KITLG
MCIP region FGF9
MCIP region CEBPA
MCIP region TLE4
MCIP region MEIS1
MCIP region LEF1

Figure S5. Integration of methylation and gene expression data reveals widespread silencing of transcription factors in CIMP. A. Box plots displaying methylation levels (MCIP-seq) computed by DiffBind at genomic regions of the Blueprint regulatory build [36]. **B.** Box plots displaying methylation levels at MCIP-seq peaks, grouped by their proximity to genomic features as determined by ChIPseeker. **C.** Average methylation levels in a 4-kb window around the center of H3K4me1 peaks, corresponding to putative promoters, for leukemia and healthy cell types. Leukemia cases, in particular CIMP and T-ALL, exhibit increased methylation levels at the center of the peak compared to differentiated cell types. **D.** Tornado plot depicting methylation levels at promoters in CIMP leukemia and healthy cells, including differentiated cell types. The HSPC tracks in purple were downloaded from ENCODE [37]and show chromatin accessibility (DNase-seq) as well as histone marks for enhancers (H3K4me1), promoters (H3K4me3), activation (H3K27ac) and repression (H3K27me3). GC density was downloaded from the UCSC browser [38]. Methylation is notably increased in CIMP leukemia, especially in regions marked by H3K27me3. **E.** Average methylation levels in a 1-kb window around the center of H3K4me1 peaks, either overlapping H3K27me3 peaks (bivalent regions, right) or not (left), for different leukemias and healthy cell types. **F.** Box plots depicting the same data as in **E.**, but only from the central 100 bp of each peak. **G.** Raw MCIP-seq data for a few selected samples of each leukemia (CIMP, T-ALL, AML) at promoters of hematopoietic genes with significant changes in methylation. **H.** Summary of results of pre-ranked GSEA conducted on genes in the vicinity of DMRs between CIMP and T-ALL, using the FDR as the ranking value. The C2 (left) and C5 (right) MSigDB collections were used in the analysis. **I.** Same as H, but showing enrichment in CIMP with respect to CD34+ cells. **J-L.** GSEA enrichment plots of differential methylation between CIMP and AML (**J**), T-ALL (**K**) and CD34+ cells (**L**). **M.** Results of motif enrichment analysis in MCIP-seq peaks conducted with the AME tool. The enrichment scores are presented in a scale from white (lowest) to red (highest). The left panel depicts motifs overrepresented in all peaks, whereas the middle and left depict motifs overrepresented in CIMP-specific peaks compared to AML and T-ALL respectively.

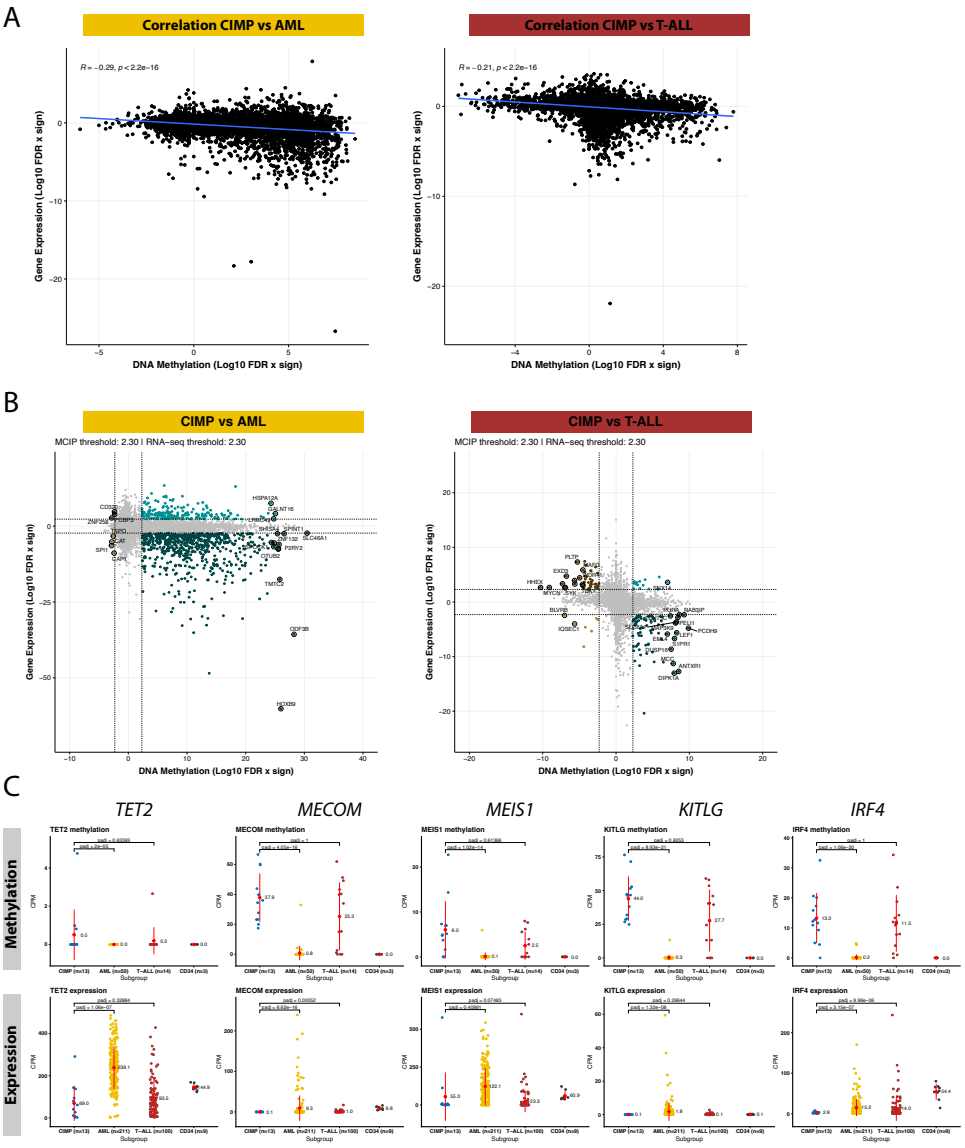


Figure S6. Integration of methylation and gene expression data. A. Scatter plot showing the inverse correlation between differences in promoter methylation (X axis) and differences in gene expression (Y axis) in CIMP compared to AML (left panel) and to T-ALL (right panel). The values are the log10 of the false discovery rate (FDR) with the sign of the fold change in comparisons by DESeq2. **B.** Starbust plot depicting changes in gene expression (Y axis) and methylation (X axis) between CIMP and AML (left) and T-ALL (right). Contrary to Figure 6, this plot includes all gene promoters, not only a selected subset. **C.** Jitter plots showing methylation (top) and expression (bottom) of a few selected genes in CIMP and other leukemias, as well as HSPCs.

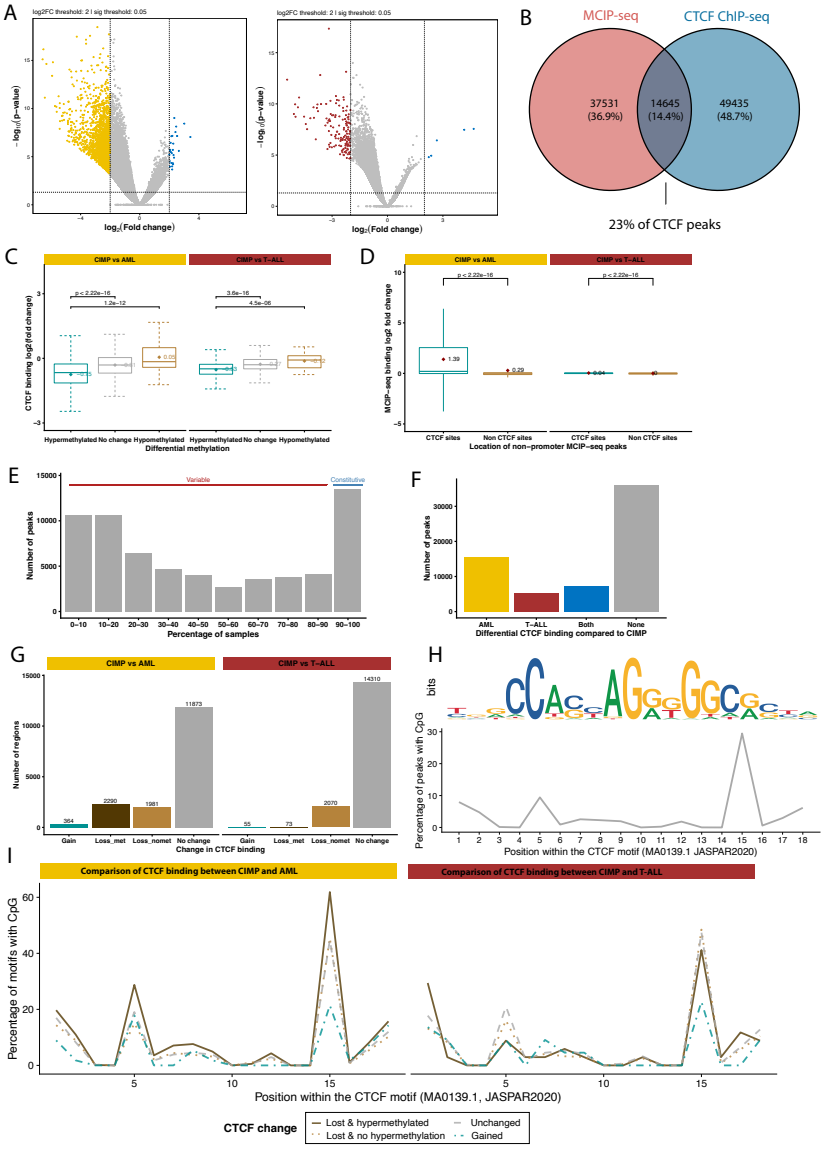
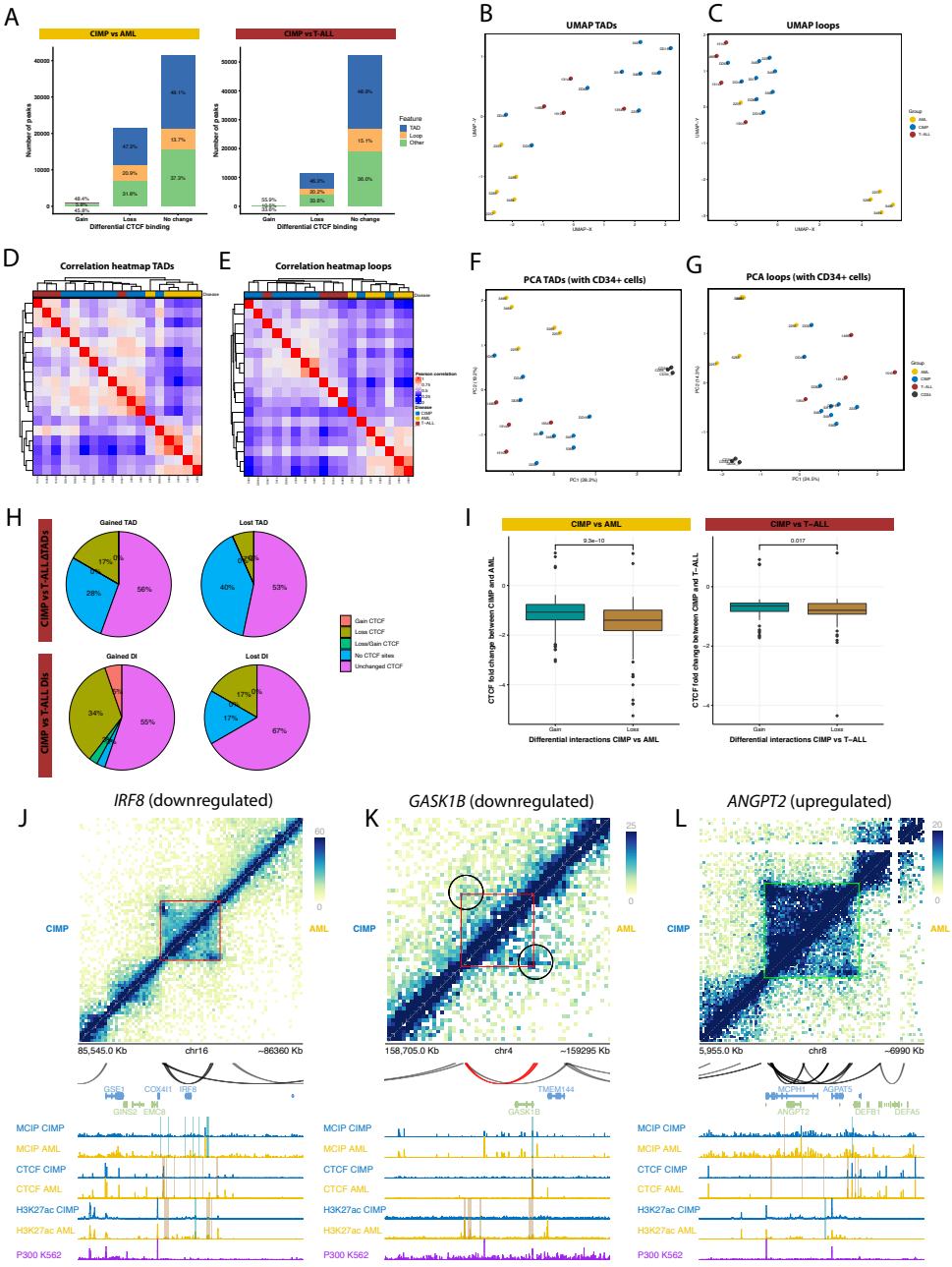


Figure S7. Effects of methylation on CTCF binding in CIMP and other leukemias. A. Volcano plot of differential CTCF binding. Regions with a FDR < 0.05 and |log2 FC| > 2 are highlighted. **B.** Venn diagram with overlap between the consensus lists of MCIP-seq peaks and CTCF ChIP-seq peaks. **C.** Box plot displaying changes in CTCF binding in relation to differences in methylation between CIMP and AML (left) or T-ALL (right). **D.** Box plot displaying changes in methylation between CIMP and AML (left) or T-ALL (right) at MCIP-seq peaks that either overlap with CTCF binding sites or that do not. **E.** Occupancy of CTCF binding sites across the entire cohort of samples, including leukemia and CD34+ cells. **F.** Number of variable CTCF peaks in comparisons between CIMP and AML or T-ALL. **G.** Number of CTCF binding sites that are gained (Gain), lost with hypermethylation (Lost_met), lost without change in methylation (Loss_nomet) or unchanged in comparisons between CIMP and AML. Only regions with data from both -CTCF ChIP-seq and MCIP-seq are considered. **H.** Average frequency of CpG dinucleotides in all detected CTCF binding sites at every position of the CTCF motif (MA0139.1, JASPAR database[39]). **I**. Same as H., but average CpG frequencies are calculated for the fractions of differential CTCF peaks described in **F**.


A
CIMP vs AML
CIMP vs T-ALL
Number of peaks
Differential CTCF binding
Gain
Loss
No change
Feature
TAD
Loop
Other
B
UMAP TADs
C
UMAP loops
D
Correlation heatmap TADs
E
Correlation heatmap loops
F
PCA TADs (with CD34+ cells)
G
PCA loops (with CD34+ cells)
H
Gained TAD
Lost TAD
Gained DI
Lost DI
CIMP vs T-ALL ΔTADs
CIMP vs T-ALL DIs
Gain CTCF
Loss CTCF
Loss/Gain CTCF
No CTCF sites
Unchanged CTCF
I
CIMP vs AML
CIMP vs T-ALL
CTCF fold change between CIMP and AML
CTCF fold change between CIMP and T-ALL
Differential interactions CIMP vs AML
Differential interactions CIMP vs T-ALL
J
IRF8 (downregulated)
K
GASK1B (downregulated)
L
ANGPT2 (upregulated)
CIMP
AML
MCIP CIMP
MCIP AML
CTCF CIMP
CTCF AML
H3K27ac CIMP
H3K27ac AML
P300 K562

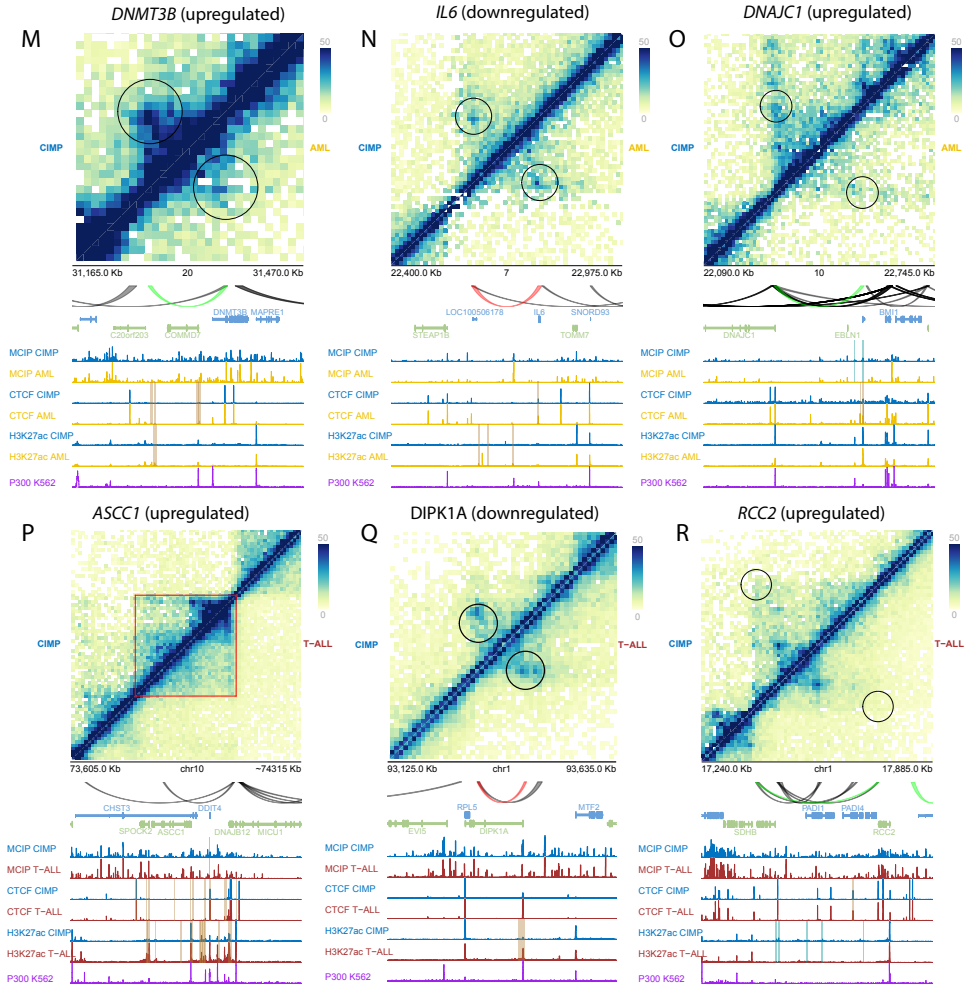


Figure S8. Differences in 3D genome organization between leukemia groups. A. Bar plots depicting the percentage of differential CTCF peaks that overlap with TAD boundaries or loop anchors. **B.** UMAP plot of TAD inclusion ratios (IR) calculated by HOMER in Hi-C data from leukemia samples. **C.** UMAP plot of loop density scores calculated by HOMER in Hi-C data from leukemia samples. **D**. Pearson correlation heatmap of TAD IRs calculated by HOMER. **E.** Pearson correlation heatmap of loop density scores calculated by HOMER. **F.** PCA plot of TAD IRs calculated by HOMER in Hi-C data from leukemias and healthy CD34+ cells. **G.** PCA plot of loop density scores calculated by HOMER in Hi-C data from leukemia and healthy CD34+ cells. **H.** Distribution of gains or losses in CTCF binding in variable TADs (top) or differential interaction (bottom) when comparing CIMP vs T-ALL. **I.** Box plot showing change in CTCF binding (expressed as log2) at gained or lost differential interactions between CIMP and AML (left) or T-ALL (right). **J.** Aggregated HIC heatmap of the *IRF8* locus, comparing interactions between the CIMP (uppermost triangle, n=5) and AML groups (bottom triangle, n=5). ΔTADs are highlighted as squares, colored in green if insulation is gained or in red if insulation is lost; DIs are indicated with black circles. Underneath, all loops detected in this region are shown in black, if they are invariable across conditions, and in green or red if they are gained or lost in CIMP relative to AML, respectively. The tracks below display MCIP-seq, CTCF ChIP-seq and H3K27ac ChIP-seq from CIMP and AML (n=4). Peaks gained in CIMP are highlighted in turquoise, whereas lost peaks are highlighted in light brown. The last track shows p300 binding measured by ChIP-seq in the K562 cell line. **K-O.** Same as H., but the *GASK1B*, *ANGPT2*, *DNMT3B*, *IL6* and loci are shown, respectively. **P-R.** Same as H., but for comparisons between CIMP (n=4) and T-ALL (n=4). The *ASCC1*, *DIPK1A* and *RCC2* loci are shown, respectively.

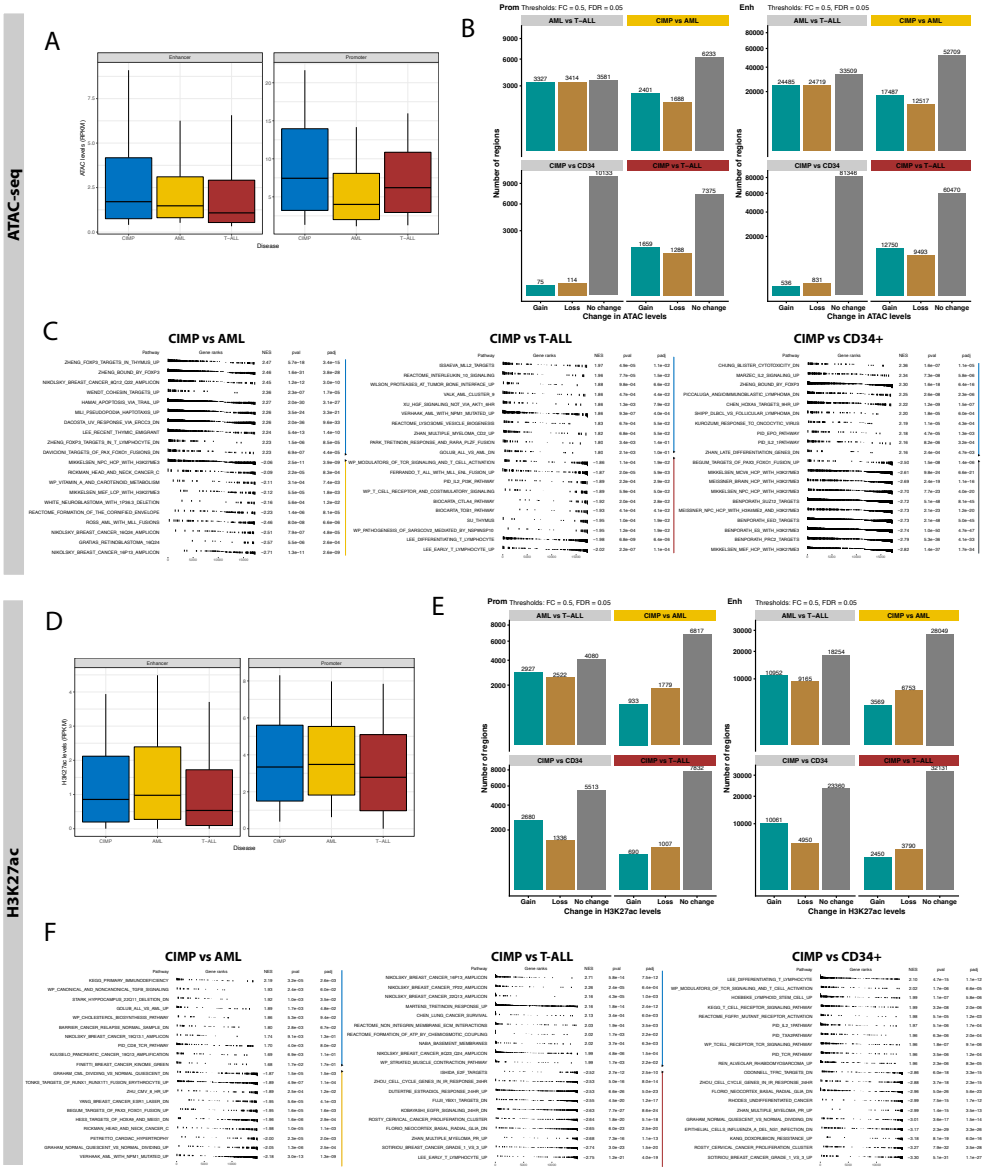
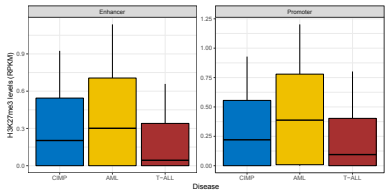


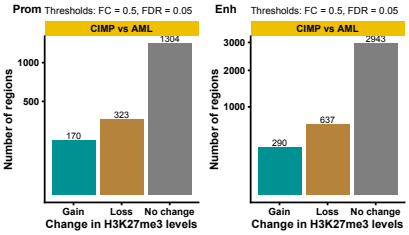
Figure S9. Additional analyses of epigenomics data. A. Box plots displaying open chromatin levels (ATAC-seq) computed by DiffBind at enhancers and promoters. **B.** Bar plot of differentially accessible regions in various supervised comparisons. A threshold of FDR < 0.05 and |log2 FC| > 0.5 was used to determine significance. **C.** Bar plot showing the top results (10 highest and 10 lowest) from gene set enrichment analysis (GSEA) conducted on genes close to open chromatin peaks, using the C2 collection. Ranking of genes was based on differential chromatin accessibility for each comparison, i.e. CIMP vs AML (left), CIMP vs T-ALL (middle) and CIMP vs CD34+ cells (right). **D-F.** Same as A-C, but using H3K27ac ChIP-seq data instead. **G-I.** Same as A-C, but using H3K27me3 data instead.

H3K27me3

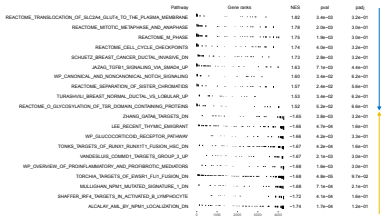
G



H



I



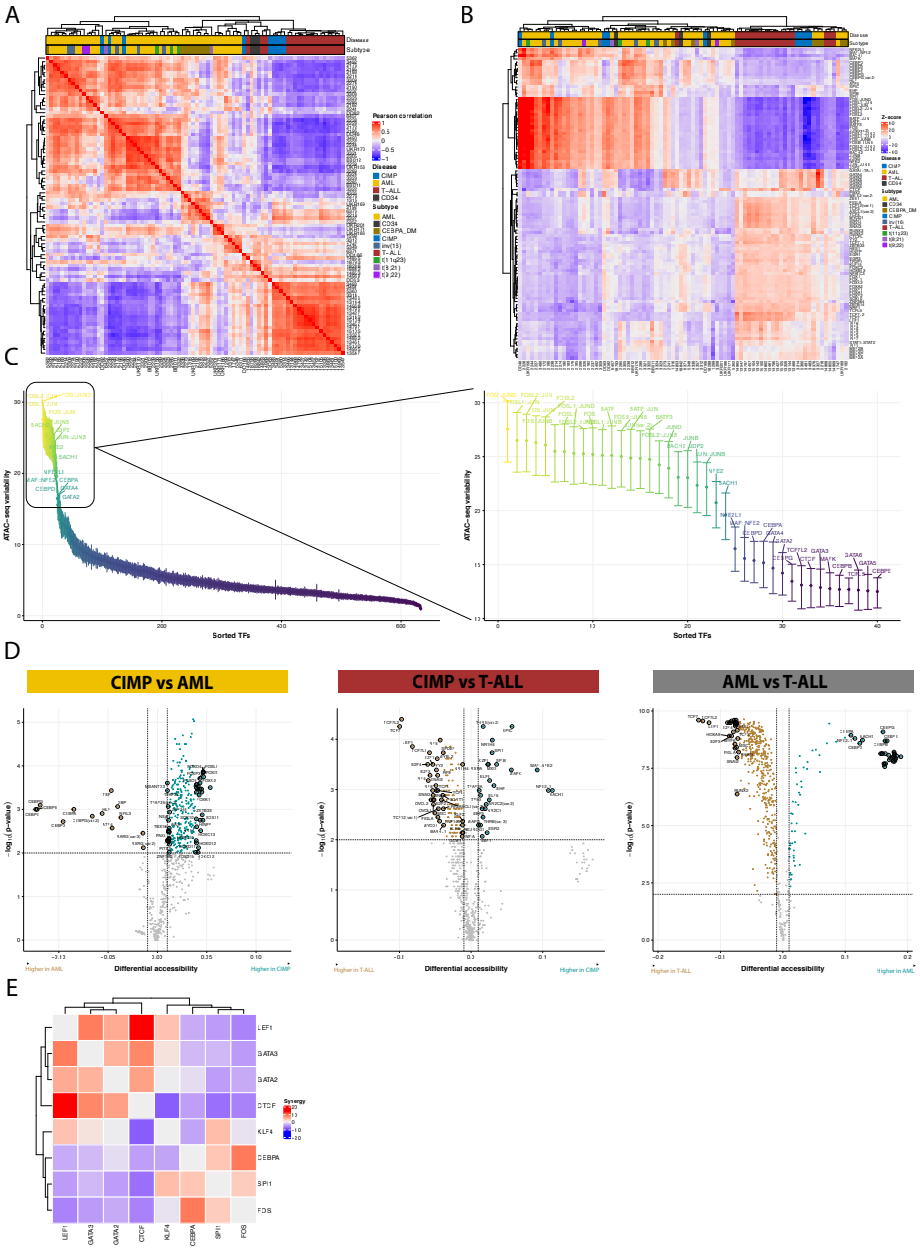


Figure S10. Estimation of motif activity based on open chromatin data. A. Heatmap of Pearson correlation between the motif activities of each sample. **B.** Heatmap displaying the motif activities, as Z-scores, of the 100 most variable motifs across all samples. Hierarchical clustering was preformed based on the Euclidean distance. **C.** Transcription factors motifs ranked by variability in chromatin accessibility, as measured by ATAC-seq. The top 40 are shown as a zoom-in on the right. **D.** Volcano plots displaying comparisons in motif activity between CIMP and AML (left), CIMP and T-ALL (middle), and AML and T-ALL (right). Motifs with a p-value < 0.01 (Wilcoxon signed-rank test) and |differential deviation| > 0.01 are highlighted. **E.** Heatmap depicting the synergy between a subset of variable TFs, defined as the deviation of chromatin accessibility in peaks with both motifs relative to peaks with only one motif. High synergy score can indicate cooperativity or competition between TFs.

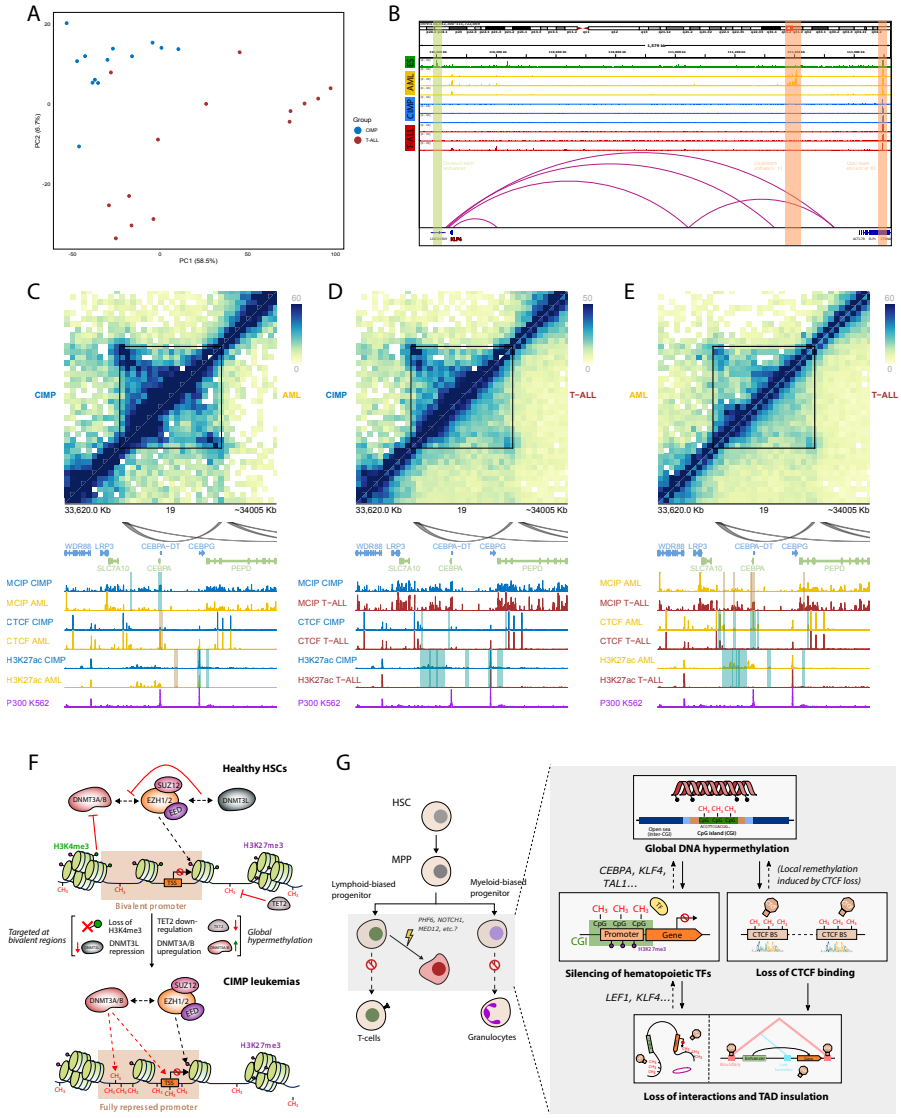


Figure S11. Additional figures for discussion. A. Principal component analysis of MCIP-seq data from CIMP and T-ALL samples **B.** Overview of the KLF4 locus displaying H3K27ac levels for embryonic stem cells (ES), AML, CIMP and T-ALL in green, yellow, blue and red respectively. Chromatin loops detected in leukemia are shown in purple below and putative enhancer regions relevant for KLF4 expression are highlighted in green (proximal) or orange (distal). **C-E.** Aggregated HIC heatmaps of the *CEBPA* locus, comparing interactions between CIMP vs AML (C), CIMP vs T-ALL (D), AML vs T-ALL **E.** The *CEBPA* TAD is indicated as a black square. Underneath, all loops detected in this region are shown in black, if they are invariable across conditions, and in green or red if they are gained or lost. The tracks below display aggregated MCIP-seq, CTCF ChIP-seq and H3K27ac ChIP-seq data (n=4 each). Peaks gained are highlighted in turquoise, whereas lost peaks are highlighted in light brown. The last track shows p300 binding measured by ChIP-seq in the K562 cell line. **F.** Proposed mechanism of preferential hypermethylation at H3K27ac-marked regions. In CIMP leukemias, lack of DNMT3L and loss of H3K4me3 at bivalent regions enables the binding of DNMT3 proteins, recruited by EZH2. Moreover, lack of TET2 prevents active demethylation. **G.** Diagram summarizing the epigenetic mechanisms described in this study leading to differentiation block in CIMP leukemia.

REFERENCES

1. Tyner, J. W. *et al.* Functional genomic landscape of acute myeloid leukaemia. *Nature* **562**, 526–531 (2018).
2. The Cancer Genome Atlas Research Network *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–74 (2013).
3. Kalender Atak, Z. *et al.* Comprehensive Analysis of Transcriptome Variation Uncovers Known and Novel Driver Events in T-Cell Acute Lymphoblastic Leukemia. *PLOS Genet.* **9**, e1003997 (2013).
4. Liu, Y. *et al.* The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. *Nat. Genet.* **49**, 1211–1218 (2017).
5. Chen, B. *et al.* Identification of fusion genes and characterization of transcriptome features in T-cell acute lymphoblastic leukemia. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 373–378 (2017).
6. Neumann, M. *et al.* Mutational spectrum of adult T-ALL. *Oncotarget* **6**, 2754 (2015).
7. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–31 (2012).
8. Zhang, J. *et al.* The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* **481**, 157–163 (2012).
9. Neumann, M. *et al.* Whole-exome sequencing in adult ETP-ALL reveals a high rate of DNMT3A mutations. *Blood* **121**, 4749–4752 (2013).
10. Alexander, T. B. *et al.* The genetic basis and cell of origin of mixed phenotype acute leukaemia. *Nature* **562**, 373–406 (2018).
11. Xiao, W. *et al.* PHF6 and DNMT3A mutations are enriched in distinct subgroups of mixed phenotype acute leukemia with T-lineage differentiation. *Blood Adv.* **2**, 3526–3539 (2018).
12. Yang, F. *et al.* Identification and prioritization of myeloid malignancy germline variants in a large cohort of adult patients with AML. *Blood* **139**, 1208–1221 (2022).
13. Janiszewska, H. *et al.* Constitutional mutations of the CHEK2 gene are a risk factor for MDS, but not for de novo AML. *Leuk. Res.* **70**, 74–78 (2018).
14. Zhang, S. J. *et al.* Gain-of-function mutation of GATA-2 in acute myeloid transformation of chronic myeloid leukemia. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 2076–2081 (2008).
15. Churpek, J. E. *et al.* Inherited mutations in cancer susceptibility genes are common among survivors of breast cancer who develop therapy-related leukemia. *Cancer* **122**, 304–311 (2016).
16. Simonetti, G. *et al.* Aneuploid acute myeloid leukemia exhibits a signature of genomic alterations in the cell cycle and protein degradation machinery. *Cancer* **125**, 712–725 (2019).
17. Mrózek, K., Heerema, N. A. & Bloomfield, C. D. Cytogenetics in acute leukemia. *Blood Rev.* **18**, 115–136 (2004).
18. Walter, M. J. *et al.* Acquired copy number alterations in adult acute myeloid leukemia genomes. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 12950–12955 (2009).
19. Lord, K. A., Abdollahi, A., Hoffman-Liebermann, B. & Liebermann, D. A. Proto-oncogenes of the fos/jun family of transcription factors are positive regulators of myeloid differentiation. *Mol. Cell. Biol.* **13**, 841–851 (1993).
20. Elsässer, A. *et al.* The fusion protein AML1-ETO in acute myeloid leukemia with translocation t(8;21) induces c-jun protein expression via the proximal AP-1 site of the c-jun promoter in an indirect, JNK-dependent manner. *Oncogene* **22**, 5646–5657 (2003).
21. Staber, P. B. *et al.* Common alterations in gene expression and increased proliferation in recurrent acute myeloid leukemia. *Oncogene 2004 234* **23**, 894–904 (2004).

22. Rodrigues, N. P. *et al.* GATA-2 regulates granulocyte-macrophage progenitor cell function. *Blood* **112**, 4862–4873 (2008).
23. Zhang, P. *et al.* Enhancement of hematopoietic stem cell repopulating capacity and self-renewal in the absence of the transcription factor C/EBPα. *Immunity* **21**, 853–863 (2004).
24. Di Giammartino, D. C. *et al.* KLF4 is involved in the organization and regulation of pluripotency-associated 3D enhancer networks. *Nat. Cell Biol.* **21**, 1179 (2019).
25. Wei, Z. *et al.* Klf4 Organizes Long-Range Chromosomal Interactions with the Oct4 Locus in Reprogramming and Pluripotency. *Cell Stem Cell* **13**, 36–47 (2013).
26. Shan, Q. *et al.* Tcf1 and Lef1 provide constant supervision to mature CD8+ T cell identity and function by organizing genomic architecture. *Nat. Commun. 2021 121* **12**, 1–20 (2021).
27. Cai, D. H. *et al.* C/EBP alpha:AP-1 leucine zipper heterodimers bind novel DNA elements, activate the PU.1 promoter and direct monocyte lineage commitment more potently than C/EBP alpha homodimers or AP-1. *Oncogene* **27**, 2772–2779 (2008).
28. Rangatia, J. *et al.* Elevated c-Jun expression in acute myeloid leukemias inhibits C/EBPα DNA binding via leucine zipper domain interaction. *Oncogene* **22**, 4760–4764 (2003).
29. Nerlov, C. & Graf, T. PU.1 induces myeloid lineage commitment in multipotent hematopoietic progenitors. *Genes Dev.* **12**, 2403–2412 (1998).
30. Anderson, M. K., Weiss, A. H., Hernandez-Hoyos, G., Dionne, C. J. & Rothenberg, E. V. Constitutive expression of PU.1 in fetal hematopoietic progenitors blocks T cell development at the pro-T cell stage. *Immunity* **16**, 285–296 (2002).
31. Itoh-Nakadai, A. *et al.* The transcription repressors Bach2 and Bach1 promote B cell development by repressing the myeloid program. *Nat. Immunol.* **15**, 1171–1180 (2014).
32. Novershtern, N. *et al.* Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell* **144**, 296–309 (2011).
33. Laurenti, E. *et al.* The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment. *Nat. Immunol.* **14**, 756–763 (2013).
34. Xie, X. *et al.* Single-cell transcriptomic landscape of human blood cells. *Natl. Sci. Rev.* **8**, (2021).
35. Hay, S. B., Ferchen, K., Chetal, K., Grimes, H. L. & Salomonis, N. The Human Cell Atlas bone marrow single-cell interactive web portal. *Exp. Hematol.* **68**, 51–61 (2018).
36. Martens, J. H. A. & Stunnenberg, H. G. BLUEPRINT: mapping human blood cell epigenomes. *Haematologica* **98**, 1487–9 (2013).
37. Feingold, E. A. *et al.* The ENCODE (ENCyclopedia of DNA Elements) Project. *Science* vol. 306 636–640 (2004).
38. Speir, M. L. *et al.* The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.* **44**, D717-25 (2016).
39. Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**, D91-4 (2004).

CHAPTER 8

Summary and general discussion

1. SUMMARY OF FINDINGS

Every day, the hematopoietic system produces billions of blood cells through the progressive specialization of hematopoietic stem cells (HSCs), a process known as hematopoiesis. Epigenetic mechanisms lie at the heart of this process, precisely establishing the transcriptional program of blood cells along the differentiation continuum. Disruptions in these mechanisms can give rise to dysregulation of genes critically involved in proliferation or differentiation, ultimately resulting in leukemia. In fact, almost 75% of AML patients carry mutations in epigenetic modifiers. The concept of epigenetics as a bridge between genotype and phenotype was originally defined in the 1940s by Conrad Waddington, who metaphorically depicted the epigenetic landscape as hilly slope that guides cell differentiation. Our understanding of epigenetics has come a long way since those early days, and the valleys and ridges in Waddington's model are no longer seen as theoretical constructs, but measurable features such as DNA methylation, histone acetylation and transcription factor binding. The addition of next generation sequencing (NGS) technologies to the arsenal of the molecular biologist has made it possible to chart the epigenetic landscape of entire genomes, even in single cells. Nevertheless, this unprecedented wealth of data has come with its own unique set of challenges, namely in terms of data analysis and interpretation – bioinformatics has emerged as an indispensable discipline to face these challenges.

The work presented in this thesis draws from molecular biology and bioinformatics alike to shed light on the role of transcription factors (TF) in healthy hematopoiesis, explore the mechanisms of enhancer hijacking in leukemia and identify novel epigenetic events underlying leukemogenesis.

In **chapter 2**, we investigated the functional requirement for C/EBPA in myeloid differentiation and HSC maintenance in a mouse model lacking the hematopoietic +37 *Cebpa* enhancer (+42 kb in humans). It had been previously observed that deletion of the *Cebpa* gene [1] or its hematopoietic enhancer [2,3] leads to neutropenia with concomitant depletion of the LT-HSC compartment. Although prior reports concluded that the latter was a consequence of a direct role of C/EBPA in LT-HSC function, here we provide evidence indicating it is a cell-extrinsic event triggered by neutropenia. This conclusion was supported by four lines of evidence: *Cebpa* was barely detectable in LT-HSCs using single cell transcriptomics, LT-HSC loss was proportional to the degree of neutropenia, lymphocytes were normally produced in mice lacking *Cebpa*, and secondary transplants of 37kb-deleted bone marrow also led to neutropenia.

The work in **chapters 3 to 5** concerns enhancer hijacking in AML with 3q26 rearrangements. Repositioning of the *GATA2* hematopoietic super-enhancer (SE) to the *MECOM* locus on 3q26 drives overexpression of the *EVI1* isoform in AML with inv(3)/t(3;3), accompanied by loss of *GATA2* expression [4,5]. The *MDS1-EVI1* transcriptional isoform, encoded by *MECOM*

as well, is not expressed in these leukemias. In **chapter 3**, we demonstrated that a similar mechanism operates in AMLs with atypical 3q26 rearrangements, all of which bring a SE into the vicinity of *EVI1*, resulting in its overexpression. Furthermore, they also lack *MDS1-EVI1* expression and frequently exhibit monoallelic expression of *GATA2* even though it is not directly implicated in the translocation. Altogether, we concluded that AMLs with 3q26 rearrangements constitute a single disease entity with common pathobiological features. **Chapter 4** explores the mechanism of SE hijacking by *EVI1* in AML with t(3;8), which repositions a *MYC* SE to *EVI1*. We created a human-based, K562-derived cell line with t(3;8) by applying CRISPR-Cas9 technology directed at the breakpoints identified in a t(3;8) AML patient by 3q-capture. This model carried e*GFP* downstream of *EVI1* (*EVI1-eGFP*), allowing us to track changes in *EVI1* expression upon deletion of functional regions. This strategy identified a hematopoietic module of the *MYC* SE critical for *EVI1* expression and enhancer-promoter interaction. This interaction was dependent on the binding of CTCF both to the *MYC* SE and to the promoter of *EVI1* in convergent orientation. Characterization of other 3q26-rearranged AMLs showed the CTCF upstream of *EVI1* is always preserved, suggesting this site is essential for enhancer hijacking. In **chapter 5**, we identified motifs in the rearranged *GATA2* SE that are required for aberrant *EVI1* expression. To this end, we conducted a CRISPR/Cas9-based scan of a minimally translocated region of the *GATA2* SE in MUTZ3 with *EVI1*-eGFP, using a lentiviral library of 3,239 sgRNAs. Cells lacking GFP/EVI1 expression were enriched for sgRNAs targeting a p300-interacting site bound by a heptad of hematopoietic TFs, and particularly a MYB-binding site within this region. Mutations of this site led to a reduction in *EVI1* without affecting *GATA2* transcription. Pharmacological inhibition of MYB achieved the same outcome, suggesting a therapeutic avenue to selectively interfere with oncogenic activation of *EVI1*.

The studies in chapters 6 and 7 explore other mechanisms of epigenetic dysregulation in AML and their role in leukemogenesis. In **chapter 6**, we integrated whole exome sequencing (WES) and RNA-seq data from 200 AMLs to measure allele-specific expression (ASE), which acts as a surrogate marker for changes in *cis*-regulatory regions. This unbiased analysis detected frequent ASE of *GATA2* in AML and, particularly, in 95% of the patients with double *CEBPA* mutations (*CEBPA* DM). As others have reported [6], many of those *CEBPA* DM AMLs also exhibited *GATA2* mutations, in which case the mutated allele was always preferentially expressed. We further established that *GATA2* ASE is a somatic event absent in remission and that it stems from allele-specific methylaton of the promoter with concomitant hyperactivation of the enhancer in the other allele. This phenomenon is possibly an example of primary epimutation that cooperates with a genetic hit to drive leukemogenesis. The impact of methylation in the regulation of myeloid TFs was further examined in **chapter 7**. Previous work from Wouters et al. had identified a subgroup of leukemias defined by widespread methylation leading to silencing of *CEBPA* and a mixed myeloid/lymphoid phenotype [7,8]. Separately, Gebhard and colleagues found a similar subgroup with a CpG island methylation phenotype (CIMP)[9]. In this chapter, we showed that CIMP and *CEBPA*-

silenced leukemias are a distinct entity very similar to ETP-ALL, characterized by a shared epigenetic signature rather than a common set of gene mutations. These leukemias exhibit methylation patterns similar to T-ALL, but with epigenetic and transcriptional profiles intermediate between AML and T-ALL. They are likely to derive from an early progenitor in which promoter methylation at myeloid TFs aborts myeloid differentiation. Furthermore, their hypermethylation also leads to loss of CTCF binding at sites with CpGs, inducing changes in genome structure and dysregulation of nearby genes.

In closing, this thesis confirms the importance of strict epigenetic regulation of hematopoiesis, with alterations in the relevant mechanisms promoting bone marrow failure or leukemia. In particular, enhancer hijacking and aberrant promoter methylation, among others, lead to aberrant expression of critical hematopoietic regulators such as *CEBPA*, *EVI1* or *GATA2* in AML. Knowledge of these epigenetic alterations and their molecular underpinnings opens the door to the development of targeted therapies that could selectively reverse the leukemogenic process.

2. GENERAL DISCUSSION

2.1 Transcription factors in healthy hematopoiesis: the role of *CEBPA* in HSC maintenance

CCAAT-enhancer binding protein alpha (C/EBPA), encoded by the *CEBPA* gene, is a master regulator of myelopoiesis, required for GMP formation [10,11], granulopoiesis [12,13] and monopoiesis [14]. At the GMP stage, C/EBPA acts as a regulatory switch: while high levels can set off granulocyte differentiation by activating genes such as *GFI1* or *CEBPE,* low levels direct monopoiesis [15,16]. Besides, C/EBPA suppresses cell proliferation to shift the balance towards terminal differentiation [17–19]. Like all other members of the C/EBP family, C/EBPA is a TF with a basic leucine zipper (bZIP) domain for DNA binding and dimerization located in the C-terminal region [20]. In addition, C/EBPA is equipped with two N-terminal transactivation domains that stimulate the transcription of target genes, including key myeloid genes like those encoding the receptors for M-CSF (*CSF1R*) [21], GM-CSF (*CSF2R*) [22] and G-CSF (*CSF3R*) [23]. The expression of *CEBPA* in myeloid cells is primarily driven by interaction of its promoter with an enhancer located +42 kb downstream (+37 kb in mouse), which is uniquely active in blood tissues [2,24]. Several other putative enhancers can be identified by H3K27ac in the vicinity of *CEBPA*, but only the +42 kb and +9 kb enhancers are accessible in HSPCs. Since hematopoietic TFs bind only to the +42 kb enhancer, it probably initiates *CEBPA* expression in HSPCs, enabling the transition from CMP to GMP [2].

In keeping with the above, disruption of either *Cebpa* [10–12] or its hematopoietic enhancer [2,3] in mice blocks differentiation from CMP to GMP, resulting in a complete loss of granulocytes. Intriguingly, **several of these studies revealed depletion of the LT-HSC compartment upon loss of *Cebpa*** [1–3], whereas others reported the opposite effect [11,25]. The increased number of LT-HSCs in the latter was attributed to enhanced proliferation in the absence of the cell cycle

suppressing function of C/EBPA, as had previously been shown [26]. In the long term, however, excessive proliferation could lead to stem cell exhaustion and loss of LT-HSCs, explaining the discrepancy between experiments [1]. Either way, these conclusions imply a direct role for C/EBPA in the regulation of LT-HSC self-renewal. However, the short term boost in LT-HSC numbers and its subsequent depletion could be instead an indirect consequence of exacerbated demand for additional progenitors via a feedback mechanism [2,3].

In **chapter 2**, we set out to distinguish between these two possibilities: is the loss of LT-HSCs in *Cebpa*-null models caused by a cell-intrinsic mechanism or is it an indirect consequence of neutropenia? To this end, we employed mice with a heterozygous (37kb^{HET}) or homozygous (37kb^{HOM}) deletion of the +37 hematopoietic enhancer. Multiple lines of evidence led us to conclude that *Cebpa* loss does not directly affect LT-HSC function, and it is the loss of committed progenitors that indirectly leads to depletion in the LT-HSC compartment (Figure 1).

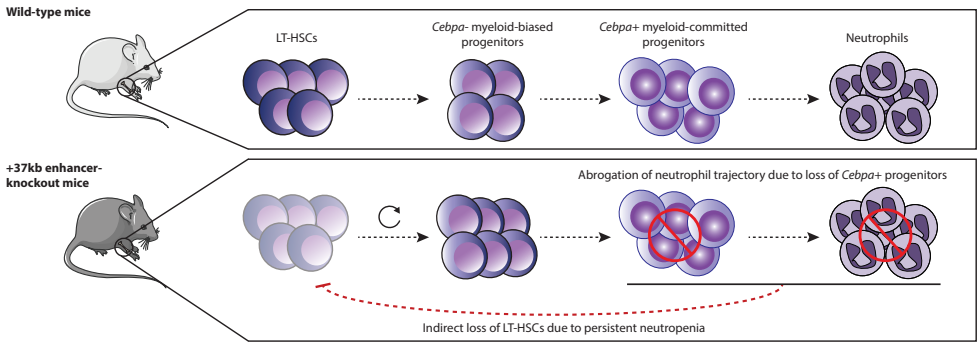


Figure 1. LT-HSC depletion is an indirect consequence of persistent neutropenia.

2.1.1 Is CEBPA present in LT-HSCs? Insight from single cell transcriptomics

Expression of *Cebpa* in HSCs is an indispensable requirement for any potential function in those cells, such as the regulation of self-renewal proposed by Hasemann and others [1]. Early analyses with RT-PCR detected low transcript levels of *Cebpa* in murine HSCs defined by the immunophenotype Lin$^-$ Sca-1$^-$ c-Kit$^+$ (LSK) IL-7Rα$^-$ [11,27–29]. Data from studies employing microarrays or RNA-seq in various progenitor populations support similar conclusions in mice [30,31] and humans [32]. Transgenic mice with an EYFP reporter exhibited *Cebpa* expression in the majority of LSK cells, but only 4% of LT-HSCs with signaling lymphocyte activation molecule (SLAM) markers, suggesting that *Cebpa* expression is mainly restricted to MPPs within the HSPC compartment [33]. Another *in vivo* study with an hCD4 reporter behind a construct consisting of the +37kb enhancer plus the *Cebpa* promoter detected signal in 17% of SLAM LT-HSC cells [24], but such a model may not fully recapitulate the native context of the *Cebpa* locus.

A shortcoming of most of these studies is that they were **conducted using bulk sequencing and RT-PCR in cell fractions isolated on the basis of immunophenotypic markers**. Early work of Irving Weissman established that HSCs are enriched in populations with specific markers [34,35], but this should not be equated with functional identity. Ultimately, LT-HSCs are defined by their capacity to reconstitute an entire blood system [36], which can only be determined by functional assays. Although the association between phenotype and HSCs is strong enough to justify the use of immunophenotype as a surrogate marker in the absence of functional assays, differences between subpopulations may be averaged out in such studies. One cannot exclude the possibility that a few mature cells can retain surface markers typically assigned to a progenitor. Moreover, the gating thresholds that discriminate different cell populations in sorting procedures are a largely arbitrary attempt to segregate parts of a continuous differentiation gradient. To complicate matters further, different combinations of markers have been proposed to delineate LT-HSCs. In the same study, 9% of LT-HSCs defined as CD34– LSK exhibited *Cebpa* expression, but only 4% of those with SLAM markers [33]. In some cases, the markers employed are inadequate to discriminate LT-HSCs from other early HSPCs.

In contrast, single cell transcriptomics offers an unbiased window into the heterogeneity of the bone marrow, including the HSC compartment. Using previously published single cell datasets, **we determined that *Cebpa* expression is virtually absent in primitive LT-HSCs**, characterized not only by immunophenotypic markers, but also by genes associated with quiescence and stemness (**chapter 2**). The 0.8% of LT-HSCs that did express *Cebpa* could be an artifact stemming from either contamination or misclassification. Accordingly, unsupervised clustering revealed some discrepancies between labels based on surface markers and the transcriptomic landscape of the cells.

Despite the advantages of single cell data, sensitivity can be a limiting factor. Thus, we cannot exclude the possibility that *Cebpa* transcripts are present at low levels in more than 0.8% of the LT-HSCs. At least two types of technical limitations should be considered:

a) **Failure to detect lowly expressed transcripts:** the detection limit of the Fluidigm C1 technique is 0.25 transcripts per million (TPM), which could possibly exclude cells with very low *Cebpa* expression. However, the physiological relevance of such lowly expressed transcripts is unclear. In fact, these extremely small levels of *Cebpa* could be background signal from pervasive transcription, which is well-documented in eukaryotes [37]. Although there is no straightforward equivalence between TPM and RNA content, studies often use cut-offs above 0.5 TPM to discriminate between transcriptional noise and genuine expression.

b) **Dropout events:** "drop-out" events can also occur in single cell RNA-seq if a transcript is missed during the initial reverse amplification [38]. Thus, at medium or even high transcription levels, *Cebpa* could be missed in a fraction of LT-HSCs. However, dropout rates are much smaller in the Fluidigm C1 platform than in droplet-based technologies [39,40]. In any case, the complete depletion of the LT-HSC pool would not be the result only from direct loss of *Cebpa*-expressing cells.

The evidence from single cell data and reporter-based studies suggests that the vast majority of LT-HSCs do not express *Cebpa* at physiological levels. Instead, *Cebpa* only becomes expressed in a subset of early myeloid biased-progenitors within the HSPC compartment, which could explain its detection at low levels by bulk strategies. Aside from technical aspects, low levels of *Cebpa* detected in a few cells can be due to pervasive transcription. This possibility ties in with the concept of "multilineage priming" whereby regulatory regions of lineage-specific regions are accessible even before they are actively expressed [41,42].

In light of this evidence, we propose that a) previously reported functions of *Cebpa* in HSC self-renewal should be reassessed; b) depletion of LT-HSC is not a cell-intrinsic effect of *Cebpa* deletion. Nevertheless, further research with more sensitive techniques and functional repopulation assays should be conducted to completely establish whether *Cebpa* is present at relevant physiological levels in LT-HSCs.

2.1.2 How does neutropenia lead to loss of the LT-HSC compartment?

One of the most striking findings in **chapter 2** is that wild type mice transplanted with +37kb^{HOM} bone marrow also acquired full-blown neutropenia and loss of the LT-HSC population. This observation strongly argues in favor of a model whereby neutropenia triggers systemic effects leading to LT-HSC depletion, as the recipient mice had healthy LT-HSCs with fully functional *Cebpa*. However, the fact that other neutropenia models do not exhibit similar defects in LT-HSCs [43,44] calls into question the idea that neutropenia itself is responsible for this phenotype. The loss of mature neutrophils may not need to be compensated by LT-HSCs directly, as there are several upstream progenitors capable of replicating to replenish the absent neutrophils. Instead, we propose that **it is the loss of early *Cebpa*-expressing myeloid progenitors that creates an exacerbated demand for LT-HSC proliferation**, pushing these cells out of quiescence. This hypothesis is supported by the downregulation of transcriptional programs associated with HSC quiescence detected in the +37kb^{HOM} mice. However, more direct proof could be obtained by conducting single cell RNA-seq on +37kb^{HOM} mice and WT controls in order to elucidate what bone marrow populations are missing and what markers define them.

Early lymphoid [45] and myeloid [46] progenitors do not occupy the same niches as HSCs. Imaging studies have shown that MPPs, which are typically considered the immediate progeny of HSCs, might be located in a close, yet physically separate niche [47,48]. We hypothesize that early *Cebpa*-expressing cells reside in close proximity to HSCs rather than in distinct myeloid niches, which could lead to disturbances in the HSC niche upon loss of these cells. Confirmation could be obtained by tracking *Cebpa* expression in spatial transcriptomics data from the bone marrow [49]. Alternatively, it may be possible to conduct live-imaging studies in mice with a fluorescent marker behind *Cebpa*. However, studies of bone marrow architecture are limited by the availability of HSC markers and the difficulty to use light microscopy on calcified tissue without disrupting it [50].

The exact nature of the feedback mechanism connecting the loss of myeloid progenitors and LT-HSC depletion remains uncertain, especially when induced in recipient mice. **How do LT-HSCs sense the loss of these progenitors?** In a normal situation, LT-HSCs remain quiescent to minimize exhaustion and cell cycle-associated DNA damage, while ST-HSCs and downstream progenitors sustain steady-state hematopoiesis [51]. However, LT-HSCs can become activated to proliferate and differentiate in response to stress, such as a serious infection or blood loss [52]. This is only possible thanks to the modulation of signaling receptors expressed on the HSC surface, which can be grouped in the following categories: a) toll-like receptors (TLR) binding pathogen-associated molecular patterns (PAMPs), b) receptors of pro-inflammatory cytokines, and c) receptors involved in cell-cell interactions within the niche [53,54]. Similar pathways may drive HSCs to exit quiescence in the context of a *Cebpa*-null bone marrow, which arguably causes a situation of stress.

Persistent activation of TLRs could result from chronic infection in the absence of neutrophils, but it hardly explains the induction of bone marrow failure in secondary transplants, not observed in control animals. Therefore, the **activation of HSCs and their subsequent exhaustion is more likely to stem from indirect alterations in the niche** leading to aberrant production of cytokines or other signals. Numerous factors are involved in the preservation of quiescence in HSCs, including TGFβ, CXCL4 and CXCL12 [50]. CXCL12 production by osteoblasts can be disrupted by G-CSF, eliciting HSC mobilization [55,56]. Thus, G-CSF produced to compensate for the lack of neutrophils in +37kb^{HOM} mice could favor HSC activation. That said, even though neutropenia patients exhibit high levels of endogenous G-CSF [57] and are regularly treated with exogenous G-CSF [58], they do not present this extreme phenotype. In fact, HSCs from patients with severe congenital neutropenia (SCN) exhibit the same mutation rate as those from healthy controls, seemingly excluding accelerated proliferation [59]. Besides, it has been reported that G-CSF induces mobilization without proliferation [60] and that deletion of *Cxcl12* in osteoblasts has no effect on HSC numbers [45]. Therefore, aberrant G-CSF signaling may be one of many mechanisms contributing to the dysregulation of HSC quiescence, but not the only one. For example, myeloid cells also maintain HSC quiescence by releasing histamine, which could be potentially lost in this context [61]. It is also possible that the feedback loop between myeloid-primed progenitors and HSCs involves molecules not yet determined. Analysis of single cell data with CellPhoneDB [62] could reveal communication networks between cell populations that are lost in +37kb^{HOM} mice.

2.1.3 Effects of deleting the +37 kb Cebpa enhancer in vivo

Non-conditional *Cebpa* knockout mice die from hypoglycemia within 8 hours of birth [63]. Although this makes it possible to analyze fetal and newborn hematopoiesis [12], studies in adults require conditional knockouts. A commonly used approach relies on Cre-mediated excision of loxP-flanked *Cebpa* upon induction of *Mx1*-Cre by interferon, which can be stimulated polyinosinic:polycytidylic acid (pI:C) [11,64]. This system is not without shortcomings,

including spontaneous deletion of the floxed gene before pI:C injection, lack of tissue specificity, incomplete deletion of the gene, and direct perturbation of HSPCs by interferon [65]. The latter is of particular concern in studies that attempt to determine a possible role for *Cebpa* in the stem cell compartment and may explain some of the observations described above, in spite of the lack of expression in LT-HSCs.

The deletion of the +37 kb enhancer addresses several of these limitations by ensuring that i) the deletion is present in all the cells, ii) loss of *Cebpa* expression is specific to the hematopoietic lineage. This is based on the findings of Avellino et al. in 2016, showing that the human +42 kb enhancer is only marked by H3K27ac in blood cells, whereas other enhancers are active in the remaining *CEBPA*-expressing tissues. Nevertheless, it should be pointed out that this analysis was restricted to ChIP-seq data available from the Roadmap project, encompassing 111 tissues [66]. Consequently, we cannot rule out the possibility that this enhancer is also active in cell types that were not assayed by the Roadmap consortium, particularly in the niche. For example, data were not generated for Schwann cells or osteoclasts, let alone for rare HSC niche components like CXCL12-abundant reticular (CAR) cells. Given the impracticality of testing every single cell type in the adult bone marrow, an attractive approach to resolve this question would be to conduct single cell ATAC-seq [67] or Cut&Tag for H3K27ac [68]. Even though this strategy would not interrogate cells types outside the bone marrow or that transiently appear during development, it is unlikely those would be relevant for the LT-HSC phenotype observed in *Cebpa*-null mice.

2.1.4 Clinical implications of LT-HSC loss in the context of neutropenia induced by CEBPA dysfunction

As mentioned above, SCN patients do not develop severe LT-HSC loss or enhanced LT-HSC proliferation that could potentially lead to exhaustion. In contrast to *Cebpa*-null mice, it is thought that the differentiation block in these patients occurs rather late in the hierarchy, namely in promyelocytes [69]. Most of these patients carry mutations in *ELANE*, which encodes for neutrophil elastase, a proteolytic enzyme that is mainly involved in neutrophil-mediated cytotoxicity. On the other hand, a reduction of CD34+ HSPCs has been reported in Shwachman-Diamond syndrome, one of whose hallmarks is neutropenia [70]. However, this rare inherited disorder involves multiple organs, as it is typically caused by mutations in *SBDS* gene and other genes that participate in ribosomal biogenesis [71]. Although loss of HSPCs in this disease may be a result of increased apoptosis through the Fas pathway [72], a possible contribution of neutropenia to LT-HSC depletion should be explored. Interestingly, HSC failure in Shwachman-Diamond syndrome has been associated with loss of cell polarity, which correlates with the degree of neutropenia [73].

The conclusions derived from our study in *Cebpa*-null mice are more likely to be applicable to patients in which the differentiation block takes place at the same stage. Both somatic and familial mutations in *CEBPA* are a frequent driver of AML [74], with double

CEBPA mutations (CEBPA DM) defining a subgroup with a unique expression profile and favorable outcome [75,76]. These mutations either interfere with the ability of CEBPA to bind DNA or prevent the expression of the full-length p42 isoform [77]. Moreover, *CEBPA* is downregulated by several other oncogenic mechanisms in various AML subgroups, including hypermethylation of its promoter, mRNA instability or protein degradation [77]. Either way, loss of *CEBPA* function leads to differentiation block and neutropenia in AML, as leukemic stem cells (LSCs) displace healthy HSCs. Like their normal counterparts, LSCs also receive signals from the microenvironment that sustain their quiescent state [78]. The proposed mechanism for LT-HSC depletion does not seem to affect LSCs, either due to cell-intrinsic alterations or modifications in the niche that favor a leukemogenic environment. However, it might be interesting to examine whether the loss of healthy LT-HSCs in the context of AML is accelerated by neutropenia, as normal cells are not only outcompeted by LSCs, but also try to cope with the additional demand for myeloid-primed progenitors.

How well does the +37kb^{HOM} model recapitulate human disease? On the one hand, abolished *CEBPA* expression due to lack of an enhancer is likely to be indistinguishable from inactivation by other mechanisms. On the other hand, loss of activity of the +42 kb *CEBPA* enhancer has been reported in AML patients, as it is the target of oncoproteins which mediate its inactivation in leukemia [77]. For example, EVI1 binds the +42 kb enhancer in AML (unpublished ChIP-seq data), which explains the reduction of *CEBPA* expression seen in patients with *EVI1* overexpression. Accordingly, Perkins and colleagues have shown that the binding of EVI1 to the murine +37 Kb enhancer represses the transcription of *Cebpa* [79]. This is in line with our observation that expression of *Evi1* inversely correlates with that of *Cebpa* in bone marrow single cell data. Similarly, the fusion oncoprotein RUNX1-RUNX1T1 also binds and inactivates the +42 kb enhancer in patients with t(8;21) AML, leading to transcriptional repression of *CEBPA* [80]. Upon degradation of RUNX1-RUNX1T1, acetylation levels of the +42 kb rapidly and transcription of *CEBPA* rapidly increase.

2.2 Enhancer hijacking in AML: lessons learned from 3q26/*MECOM* rearrangements

The correct interaction between enhancers and promoters is a critical determinant of gene expression and, consequentially, cell identity [81]. Enhancer hijacking occurs when aberrant expression of a gene is driven by an enhancer that would normally control the transcription of another gene, resulting in altered patterns of gene expression that frequently contribute to tumorigenesis [82]. This is usually the consequence of structural variants (SVs) that juxtapose a gene to an active enhancer [83,84], but it can also stem from the loss of TAD boundaries that would otherwise preclude the interaction between an enhancer and a promoter [85]. Some of the earliest examples of enhancer hijacking are the activation of *MYC* [86,87] and *BCL2* [88] by a repositioned enhancer of the immunoglobulin heavy chain (IGH) locus in lymphomas with t(8;14) and t(14;18), respectively. Nevertheless, the term only came to prominence recently with the explosion of epigenetic research enabled by NGS technologies.

AML with inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2) [inv(3)/t(3;3) AML] is a distinct entity with poor prognosis recognized by the World Health Organization (WHO) classification [89,90]. Both inv(3) and t(3;3) result in the hijacking of a *GATA2* hematopoietic enhancer by *EVI1* (encoded by *MECOM*), leading to overexpression of the latter and *GATA2* haploinsufficiency [4,5]. Aside from t(3;3)/inv(3), other recurrent translocations involving the 3q26/*MECOM* locus are found in AML, such as t(2;3), t(3;6), t(3;7) or t(3;8). These cases with atypical translocations, which constitute 35% of all 3q26-rearranged AMLs, also exhibit aberrant *EVI1* expression and poor prognosis [90].

2.2.1 A common mechanism driving EVI1 expression in 3q26-rearranged AML

Integration of 3q-capture and epigenomics data revealed that the breakpoints of atypical 3q26-rearranged AMLs are invariably located in the vicinity of a super-enhancer (SE) bound by key myeloid TFs (**chapter 3**). The genes under the control of these regions are highly expressed in HSPCs, including *THADA* [t(2;3)], *ARID1B* [t(3;6)], *CDK6* [t(3;7)] or *MYC* [t(3;8)]. These observations suggest that a **defining feature of 3q26-rearranged AML is the hijacking of a SE active in HSPCs by *EVI1*,** leading to its overexpression, while the *MDS-EVI1* isoform remains silent (Figure 2). Intriguingly, allele specific expression (ASE) or copy number loss (CNL) of *GATA2* was also present in 50% of these leukemias, even though *GATA2* was not affected by any of those translocations. In addition to the translocated SEs reported in chapter 3, we have identified other SEs involved in recurrent 3q26 rearrangements in a separate cohort, such as the *ETV6* SE in t(3;21) (Table 1).

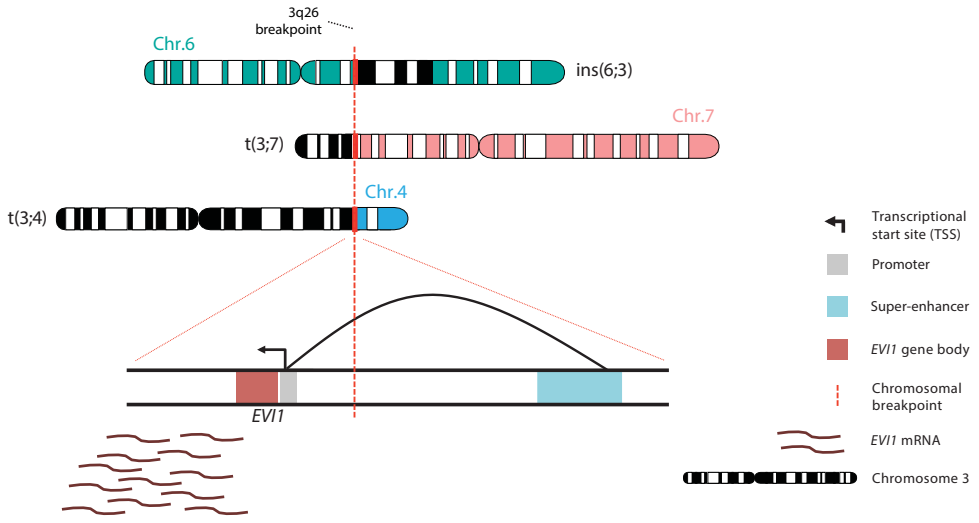


Figure 2. A common mechanism for 3q26-rearranged AML.

Taking into consideration their shared pathogenic mechanisms and their poor prognosis, we argue that AMLs with either t(3;3)/inv(3) or atypical 3q26 rearrangements belong to a single disease entity that can be generally termed "**3q26-rearranged AML**". It will be necessary to conduct further research with larger patient cohorts to ascertain whether their clinical outcome and their response to therapy are indeed comparable. Concomitant mutations and cytogenetics should be considered in this assessment, since t(3;3)/inv(3) AML co-occurs with mutations in signaling genes, splicing factors and TFs like *GATA2* [91]. The study in **chapter 3** also underscores the need for appropriate molecular assays, since some atypical 3q26 rearrangements were masked by a complex karyotype. Traditional assays like FISH and *EVI1/MDS1-EVI1* qPCR can be supplemented by 3q-capture or RNA-seq if the presence of these chromosomal aberrations is suspected.

The interaction between the rearranged super-enhancers and the promoter of *EVI1* is dependent on a CTCF binding site upstream of *EVI1*, which facilitates the creation of cohesin loops with convergently oriented CTCF sites on the enhancer side (**chapter 4**). This conclusion stems from experiments targeting CTCF binding sites in our t(3;8) K562 model, as well as the observation that 3q26 rearrangements spare the CTCF binding site upstream of *EVI1.* Future research in models recapitulating other rearrangements, like MUTZ3 or CRISPR-edited cell lines, should confirm the importance of these sites in the establishment of chromatin interactions and the regulation of *EVI1.*

2.2.2 A recurrent set of super-enhancers is involved in oncogene activation

Among the enhancers repositioned to the vicinity of *EVI1* in 3q26-rearranged AML, some are particularly recurrent (Table 1). Intriguingly, most of these regions have also been reported in translocations leading to the overexpression of the *BCL11B* gene in a subset of mixed phenotypic acute leukemia (MPAL) cases [92]. The *GATA2* SE is also thought to drive the overexpression of various *EVI1* homologs in a subset of AMLs with 3q21 rearrangements, including *PRDM16* and *PRDM1* in AML with t(1;3)(p36;q21) and t(3;6)(q21;q21) respectively [90,93,94]. Besides, hijacking of the *ETV6* SE in AML with t(12;22)(p13;q12) can elicit *MN1* overexpression with haploinsufficiency of *ETV6* [95].

This suggests the existence of features predisposing to the acquisition of certain translocations and/or their enhancers, which are likely to be shared across the regions in Table 1. One could segregate these features into three categories: a) susceptibility to DNA breakage, b) characteristics of the rearranged enhancers, c) function of the gene losing its native enhancer (Figure 3).

Table 1. Enhancer regions recurrently translocated to *EVI1*.

Translocation	Original target	Involved in *BCL11B* rearrangements
t(2;3)	*THADA*	No
t(2;3)	*BCL11A*	No
inv(3)/t(3;3)	*GATA2*	No
t(3;3)(p24;q26)	*SATB1*	Yes
t(3;6)	*ARID1B*	Yes
t(3;7)	*CDK6*	Yes
t(3;8)	*MYC*	Yes
t(3;17)	*MSI2*	No
t(3;21)	*ETV6*	Yes
t(3;21)	*NRIP1*	No
t(3;21)	*RUNX1*	Yes

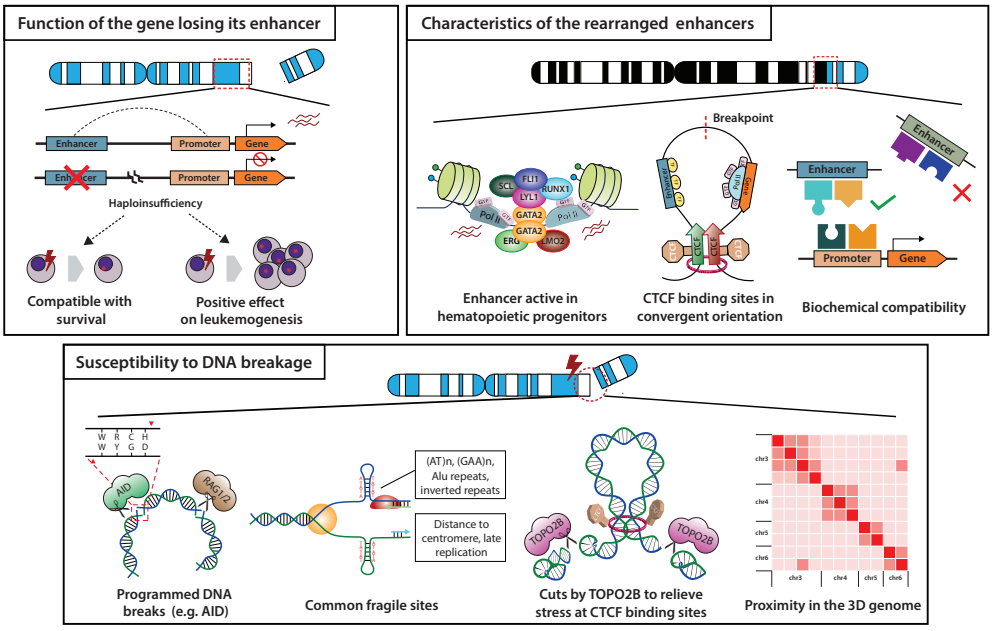


Figure 3. Proposed features driving the selectivity of enhancers hijacked by *EVI1*.

Susceptibility to DNA breakage

For a genomic rearrangement to occur, a double strand break (DSB) in the DNA must normally take place first, followed by a process of DNA repair that erroneously joins

the two ends of the translocation [96,97]. The two major pathways for DNA repair are non-homologous end joining (NHEJ) and homologous recombination (HR), both of which ensure the genomic integrity of the cell. Defects in *BRCA1* and *BRCA2*, critical components of the HR pathway, increase the likelihood of chromosome aberrations, including translocations [98]. However, these mechanisms can occasionally introduce errors even in the absence of genetic predisposition, leading to chromosomal aberrations [99]. Nicks and DSBs can occur randomly as a result of ionizing radiation or oxidative free radicals, but can also be promoted by the cellular machinery. In fact, DNA cuts are introduced by the RAG1/RAG2 and activation-induced deaminase (AID) systems as a part of V(D)J recombination and class switch recombination, respectively [97]. Untoward activity of these enzymes at off-target sites leads to recurrent translocations, such as t(11;14) (LMO2/TCR) in the case of RAG1/RAG2 or t(8;14) (MYC/IGH) caused by AID [100]. Although these systems are typically functional only in lymphopoiesis, their aberrant activation at earlier stages or the action of other enzymes with similar function could lead to DSBs at loci with specific recognition motifs.

Certain genomic regions, known as "fragile sites", are particularly prone to gaps or breaks during cell division [101]. These instable regions are hotspots for CNAs and translocations both in cultured cells and in cancer patients. Common fragile sites (CFSs) are present in all individuals, whereas a group of rare fragile sites are restricted to less than 5% of the population and are associated with a CGG-repeat expansion [101]. The fragility of CFSs is thought to result from underlying characteristics that increase the risk of failed or incomplete DNA replication [102]. A major contributor is the presence of AT-rich regions, which leads to the formation of secondary DNA structures that stall the replication fork [103]. When this happens, replication can be restarted by error-prone repair mechanisms like HR, potentially leading to rearrangements [102]. For this reason, repeats such as (AT)n, (GAA) n and (GAA)n with the potential to form secondary structures are enriched at translocation breakpoints [104]. Other determinants of genome fragility are paucity of replication origins [105], regions with late replication [106], inverted repeats [107], Alu repeats and distance to the centromere [108]. Actively transcribed long genes are more prone to rupture, possible due to collisions between replication and transcription [109,110].

Beyond the DNA sequence level, **3D spatial organization not only influences chromosome fragility, but also what regions tend to be translocation partners**. Topoisomerase 2B (TOPO2B) generates DSBs at loop anchors to prevent torsional stress during extrusion, which may lead to chromosomal aberrations upon illegitimate repair of those lesions [111,112]. Indeed, loop anchors are enriched for breakpoints of MLL-related translocations and other abnormalities [111]. On the other hand, analysis of Hi-C data in mice treated with ionizing radiations revealed that chromosomal translocations frequently involve regions that exhibit spatial proximity in the 3D genome [113].

Considering the above, it will be interesting to explore whether the breakpoints of 3q26 rearrangements are enriched in any features previously associated with chromosome

fragility, such as inverted repeats or AT-rich sequences. The requirement for CTCF-mediated loops in *EVI1* enhancer hijacking (**chapter 4**) may point to a contribution of TOPO2B-induced DSBs at loop anchors in the breakpoints of these rearrangements. In that case, the analysis of breakpoint data from sufficiently large numbers of patients should indicate whether there is enrichment for CTCF binding sites in the vicinity of the breakpoints. Binding of TOPO2B should be confirmed with ChIP-seq or Cut&Run to verify this hypothesis. Moreover, 3C-based technologies like Hi-C, HiChIP or the high-resolution MicroC [114] can reveal whether the *MECOM* locus preferentially interacts with frequent translocation partners in the 3D genome. From that perspective, it can be surmised that t(3;3) and inv(3) may be the most frequent rearrangements because interactions in *cis* are more common than in *trans*.

Features of the rearranged enhancers

Although multiple genomic sites are susceptible to DNA breakage and the formation of translocations, only a very specific set of regions are detected in recurrent 3q26 rearrangements, some of which are vastly more frequent than others. As leukemogenesis is a selective process, it stands to reason that intrinsic characteristics of the rearranged enhancers confer some form of competitive fitness. An obvious requirement is that said enhancers must be active in the cell of origin, thereby allowing the overexpression of *EVI1*, which inhibits myeloid differentiation [115] and promotes proliferation and survival of HSPCs and LSCs [116,117]. Indeed, the rearranged SEs reported in **chapter 3** exhibited H3K27ac, open chromatin and binding of hematopoietic TFs in HSPCs, based on previously published data. The genes normally under their control were also expressed in HSPCs, in line with the role of SEs in the regulation of genes involved in cell identity [118]. In sum, **a common denominator across 3q26 rearrangements is that they contain active super-enhancers in hematopoietic progenitors**. To the best of our knowledge, hijacking of these regulatory elements by genes other than *EVI1* has only been reported in blood cancers, further underscoring the cell type specificity of these oncogenic events.

Nevertheless, at least 1000 regions in the genome of HSPCs are bound by the same heptad of TFs [119] and other super-enhancers are constitutively active in most if not all tissues [120]. Why are only a few hematopoietic SEs selected, even though other SEs could potentially drive *EVI1* expression? A possible explanation is that expression levels of *EVI1* must be within a certain optimal range, thus excluding enhancers that are too strong or too weak. In keeping with this hypothesis, most cell lines do not survive upon forced overexpression of *EVI1*. Additional insight can be gleaned from research on **determinants of enhancer-promoter specificity, which include spatial architecture and biochemical compatibility** [121]. The presence of **CTCF binding sites in convergent orientation with the site at the *EVI1* promoter seems to be an important requirement for selection of an appropriate enhancer** (**chapter 4**). As outlined above, the relevance of these sites is twofold: they may increase the chances of a DSB while simultaneously allowing the interaction between the rearranged enhancer and the *EVI1* promoter. Intriguingly, additional structural proteins other than CTCF

seem to be involved in the formation and/or maintenance of chromatin loops between the *MYC* SE and *EVI1* in t(3;8). Deletion of module C within the *MYC* SE, while sparing a CTCF binding site nearby, also resulted in loss of interaction with the *EVI1* promoter. Binding of cohesin and hematopoietic TFs to that module suggests that cohesin-mediated loops could be stabilized by tissue-specific TFs, in line with previous data in other systems [122,123]. Alternatively, this role could be played by YY1, an architectural protein frequently located at loop anchors [124]. Although YY1 does not directly bind module C in K562, its presence in an adjacent region may be sufficient to stabilize loop extrusion. To tests this possibility, systematic deletion of binding sites for YY1 and hematopoietic TFs should be conducted, followed by 4C-seq to measure changes in interaction. Further experiments should be carried out in other models to establish a) what specific CTCF binding sites are essential for *EVI1* overexpression in other translocated regions, b) whether stabilization by other factors is common to all 3q26 rearrangements.

Biochemical compatibility refers to the notion that the transcriptional machinery recruited by the enhancer must be able to engage with the promoter and contribute to its activation. The sequence elements present in a core promoter determine what transcription factors and cofactors it requires; studies in Drosophila have revealed specificity of enhancers for promoters with either DPE or TATA box elements, whereas others are non-specific [125,126]. A systematic analysis in mice revealed a broad spectrum of enhancer-promoter compatibilities, with at least half of the enhancers displaying specificity for a subset of promoters [127]. Selectivity was found to be partially driven by combinations of TF motifs present in both a promoter and its enhancer. Therefore, it can be surmised that enhancers in 3q26 rearrangements recruit a number of common components that are specifically needed for the activation of the *EVI1* promoter. In **chapter 5**, a CRISPR scan in the minimally translocated region of the *GATA2* SE detected several sequences that are essential for the expression of *EVI1*, including binding sites for p300 and MYB. Pharmacological inhibition of either led to downregulation of *EVI1* in both MUTZ3 and t(3;8) K562 (data not shown), suggesting the presence of these TFs at rearranged enhancers is required for *EVI1* activation. Validation of this hypothesis should be conducted in additional models, which could be generated using the CRISPR/Cas9-based approach described in **chapter 4**.

For unbiased identification of other transcription factors and cofactors shared among all rearranged enhancers, additional CRISPR scans are possible, but they are restricted to DNA-binding proteins. This limitation could be overcome with a pull-down assay of cell lysates with biotinylated enhancer sequences as bait, followed by mass spectrometry. An alternative approach that preserves the genomic context would be to perform "reverse-ChIP" *in situ* by targeting the enhancer regions with nuclease-deactivated Cas9 (dCas9), coupled to either a FLAG tag or a biotinyl group [128,129]. Purification with either an anti-FLAG antibody or streptavidin, respectively, would allow the identification of proteins in the enhancer complex by mass spectrometry.

Function of the gene losing its native enhancer

The removal of the *GATA2* enhancer from its native locus in inv(3)/t(3;3) AML causes not only hyperactivation of *EVI1*, but also haploinsufficiency of *GATA2* [4]. Interestingly, a large number of AMLs exhibit allele-specific expression (ASE) of *GATA2* without loss of overall expression levels (**chapter 6**). GATA2 is a pivotal regulator of hematopoiesis essential for HSC generation and maintenance in both embryonic and adult stages [130–132]. Germline *GATA2* mutations leading to haploinsufficiency are associated with a number of disorders involving hematopoietic defects [133–136] and development of MDS/AML [137]. Somatic *GATA2* mutations are also found in sporadic AML, especially in cases with inv(3)/t(3;3) [91] or *CEBPA* DM [6]. In mice, *Gata2* haploinsufficiency compromises HSC proliferation and survival, as well as GMP function [138,139].

All in all, haploinsufficiency of *GATA2* in inv(3)/t(3;3) is likely to contribute to the leukemogenic process, possibly in cooperation with *EVI1*. This is corroborated by the frequent *GATA2* mutations found in these patients, in which only the mutated allele remains expressed. Furthermore, almost 50% of atypical 3q26-rearranged AMLs lose *GATA2* expression from one allele due to CNL and other unidentified mechanisms (**chapter 3**). The experimental confirmation came from *in vivo* studies showing that *Gata2* heterozygous deletions in mice with inv(3) accelerated leukemia development, owing to faster proliferation that conferred a selective advantage to *EVI1*-expressing cells [140]. Although the nature of the cooperation between *EVI1* and *GATA2* remains a conundrum, these observations support a model in which the loss of its enhancer confers a fitness advantage in inv(3)/t(3;3) AML. At the same time, the levels of *GATA2* remain elevated enough to enable HSC/LSC function, which is abolished when *GATA2* is completely lost [132]. The participation of GATA2 in other AML subtypes is further discussed in **section 2.3.2** of this discussion.

It stands to reason that similar mechanisms may operate in other 3q26 rearrangements, which also involve important hematopoietic regulators such as *MYC* [141] or *CDK6* [142]. That is**, a moderate decrease in their expression may favor AML development in the presence of *EVI1* overexpression**. Along these lines, *Myc* deficiency in mouse models shifts the balance of LT-HSCs from differentiation to self-renewal, leading to LT-HSC accumulation [141]. Even if downregulation of a certain gene does not directly contribute to leukemogenesis, partial loss of its expression should be tolerated by the affected cell. This requirement excludes translocations that would hijack enhancers from genes whose high expression levels are absolutely essential for survival. On the other hand, compensatory mechanisms may come into play at the wild type allele.

2.2.3 Unique properties of oncogenic super-enhancers offer therapeutic opportunities

One of the most striking findings about the rearranged *GATA2* enhancer in inv(3)/t(3;3) AML is that it acquires the features of a super-enhancer upon translocation, namely high levels of H3K27ac, BRD4 binding and enhancer RNA (eRNA) [4]. As a result, cell lines with 3q26 rearrangements are exquisitely sensitive to the BRD4 inhibitor JQ1, contrary to non-3q26-

rearranged models with high levels of *EVI1* expression. Despite their initial promise in mouse models and primary samples, BET inhibitors are limited by narrow therapeutic margins and underwhelming efficacy [143]. Nevertheless, they provide a useful proof of concept to develop new therapeutic strategies that exploit vulnerabilities of oncogenic super-enhancers.

One of such vulnerabilities is the dependency of the translocated *GATA2* SE on a MYB binding site that is non-functional in its native context (**chapter 5**). Deletion of this site by CRISPR/Cas9 or pharmacological inhibition of MYB resulted in selective loss of *EVI1* expression, myeloid differentiation and cell death, while *GATA2* expression remained unchanged. No changes were observed in CD34+ HSPCs from healthy donors. The CRISPR/Cas9-based scan employed in this study detected frequent mutations leading to loss of *EVI1* expression in binding sites for other TFs (TFBS), including the GATA, RUNX and MEIS families. Future research should determine whether any of them are exclusively functional in the rearranged allele, similarly to the MYB binding site. A more systematic approach to tackle this problem would be to conduct a similar CRISPR/Cas9-based scan in a t(3;3)/inv(3) cell line with different readouts for the *EVI1* and *GATA2* alleles, each of which carries a copy of the *GATA2* enhancer. We have recently developed such a model using the MUTZ3 cell line, with GFP (green) and mCherry (red) as surrogate markers for *EVI1* and *GATA2* expression respectively. Only GFP, but not mCherry, should be affected by mutations targeting TFBSs that are selectively required for *EVI1* expression. This model could also be employed in compound screens to find drug repurposing candidates that exclusively affect aberrant *EVI1* expression in cancer cells while sparing healthy cells.

Nevertheless, a fundamental enigma remains unsolved: **what factors underlie the transformation of the GATA2 hematopoietic enhancer into an oncogenic super-enhancer?** A likely explanation lies in the different regulatory elements present at each genomic context, including:

a) Novel TF binding sites: the translocations may lead to the acquisition of TFBSs in the vicinity of the breakpoint. These TFs could facilitate the binding of additional TFs at sites contained in the *GATA2* SE, such as MYB, by cooperative mechanisms like protein-protein interactions or via chromatin remodelling [144]. Formation of novel enhancers due to TFBS-creating mutations leading to the over-expression of *TAL1* [145] and *LMO2* [146] has been reported in T-ALL. Motif analysis together with ChIP-seq for suspected TFs could be used to explore this possibility.

b) Promoter switching: the characteristics of the *EVI1* promoter may be conducive to stronger activation of the rearranged enhancer; for example, via recruitment of factors that establish synergistic interactions with the TFs binding to the *GATA2* SE. This hypothesis could be tested by replacing the *EVI1* promoter with the *GATA2* promoter in the MUTZ3-eGFP cell line.

c) Loss of silencers: the activation of the SE may be elicited by the loss of cryptic transcriptional silencers present in the native locus. Given the lack of a single combination of chromatin

marks to delineate silencers [147], it may be challenging to pinpoint these elements, but screens based on H3K27me3/PRC2 coupled with 3C technologies have met some success [148].

d) Structural changes: architectural proteins such as CTCF or YY1 in a suitable orientation may favor contacts with additional enhancer modules or create favorable spatial conformations.

It is equally possible that the oncogenic super-enhancer is a result of changes in the epigenome that cannot be attributed to the underlying DNA sequence. Loss of methylation or acquisition of additional H3K27ac marks could be the result of fortuitous action by DNA demethylases (e.g. TET2) or histone acetyl transferases (e.g. p300), followed by selection of clones expressing *EVI1*. Although highly speculative, this mechanism is supported by findings in AMLs with normal karyotype in which the *GATA2* -110 kb enhancer has the properties of a super-enhancer at its native locus (**chapter 6**). Therefore, the enhancer is intrinsically capable of become a super-enhancer even in the absence of a rearrangement that alters the genomic context. This possibility could be further examined with treatments that modulate the epigenetic landscape in HSPCs, such as demethylating agents or histone deacetylase inhibitors.

Since other 3q26 rearrangements involve full-fledged super-enhancers in their native context, inv(3)/t(3;3) cell lines constitute unique models to dissect the particularities that distinguish a super-enhancer from a conventional enhancer. Any vulnerabilities unveiled by the studies proposed above should be investigated in other models, such as t(3;8), to determine whether they are present in other frequently rearranged super-enhancers and can be therapeutically exploited. Importantly, the conclusions could be applicable to other translocations involving the same SEs, such as *BCL11B* rearrangements, and potentially to unrelated instances of enhancer hijacking. It is tempting to speculate that a common set of super-enhancer features may be relevant beyond the realm of blood cancers, but many are probably tissue-specific, such as the binding of hematopoietic TFs.

2.2.4 Other mechanisms of EVI1 overexpression: secrets hidden in the non-coding genome?

It is estimated that 8% of AMLs exhibit high expression of EVI1, which is an independent predictor of poor clinical outcome [149,150]. **Only 20% of those carry 3q26 abnormalities**, where *EVI1* expression can be explained by enhancer hijacking [149]. However, this number may be an underestimation given that 3q26 rearrangements have been detected by NGS or fluorescence in situ hybridization (FISH) in cases that were missed by karyotyping. A number of these 3q26 rearrangements lead to the formation of fusion genes, such as *RUNX1-EVI1* in t(3;21)(q26;q22), in which *EVI1* overexpression is not the result of enhancer hijacking, but control by a promoter of a highly expressed gene [151].

Another 20% of those patients harbor 11q23 rearrangements involving *KMT2A*/*MLL1* [149]. It is thought that MLL fusions lead to upregulation of *EVI1* by recruiting the histone methyltransferase DOT1L to its promoter, resulting in deposition of H3K79me2, which is

associated with transcriptional activation [152–155]. Within MLL-rearranged AML, presence or absence of *EVI1* expression distinguishes two subsets of patients with differences in gene expression, morphology and immunophenotype; likewise, introduction of MLL-AF9 in murine cells only instigated expression of *EVI1* in 35% of the colonies [152]. Based on the lack of H3K79me2 at the *EVI1* promoter in committed progenitors, Bindels and colleagues hypothesized that the locus is only accessible to MLL fusions in HSCs, and thus only MLL-rearranged leukemias arising from an HSC cell of origin express *EVI1*. A subsequent study by Krivtsov and others confirmed the lack of *EVI1* expression in GMP-derived AML and further detected lower DNA methylation levels in this group [156]. Albeit outside of the scope of this thesis, it would be interesting to validate these conclusions by comparing *EVI1*-expressing cells in MLL-rearranged AML to healthy HSCs and progenitors using single-cell RNA-seq data.

On the other hand, it remains to be determined **why AMLs with MLL fusions express the longer *MDS1-EVI1* isoform**, absent in 3q26-rearranged AML [149]. Mouse studies revealed that the *Mds1-Evi1* is expressed in hemogenic endothelium during embryogenesis, but not in fetal liver [157,158]. In adult mice, *Mds1-Evi1* transcripts were found in 98% of HSCs, with LT-HSCs exhibiting much higher levels than ST-HSCs [159], as well as in most other tissues where *EVI1* is expressed [158]. However, we could not detect *MDS1-EVI1* in bulk RNA-seq data of human CD34+ HSPCs derived from healthy donors, whereas *EVI1* was expressed at moderate levels, i.e. 5-10 TPM (n=9, data not shown). Even so, it is possible that human *MDS-EVI1* is expressed in a rare progenitor subpopulation below the threshold of detection in bulk sequencing. Alternatively, chromatin at the *MDS1-EVI1* may remain open despite the lack of expression, enabling the recruitment of DOT1L by MLL fusion proteins.

A separate question, also unanswered, is **what function *MDS1-EVI1* plays in leukemogenesis**. It is believed that MDS1-EVI1 is a transcriptional activator that acts as an antagonist of EVI1, which behaves as a repressor [160]. Consistent with this notion, forced overexpression of *EVI1* in the murine myeloid line 32Dcl3 stimulated cell growth and blocked differentiation induced by G-CSF, whereas MDS-EVI1 blocked cell growth without affecting differentiation [161,162]. However, MDS1-EVI1 did not interfere with the repression of TGF-β by EVI1 in the same model, which suggests that either the isoforms are not fully antagonistic or EVI1 has a dominant role [163]. Likewise, experiments in mouse models have produced somewhat controversial results. While Zhang and others reported that *Mds1-Evi1* is essential for LT-HSC function [159], a separate study found that transduction of *MDS1-EVI1* failed to rescue a heterozygous knockout of *Mecom*, whereas *Evi1* did [164]. In light of these observations and the lack of expression in 3q26-rearranged AML, it is plausible that *MDS1-EVI1* is dispensable for leukemogenesis and its presence in MLL-rearranged AML is merely accidental. This would also explain why *MDS-EVI1* expression does not have a prognostic value [149]. The cause and possible consequences of *MDS1-EVI1* expression should be further explored in mouse models with MLL fusions, such as MLL-AF9 or MLL-AF4, which promptly develop AML [165]. Deletion of *MDS1-EVI1* in such a model could cast light on its exact contribution to pathogenesis.

Finally, the **mechanisms for *EVI1* expression remain completely unknown in almost 60% of *EVI1*-positive AML** patients, excluding possible cryptic 3q26 rearrangements [149]. In ovarian cancer, *MECOM* amplifications leads to increased expression of both *EVI1* and *MDS1-EVI1* isoforms, which is paradoxically associated with favorable prognosis in that disease [166]. Although less common, amplifications of *EVI1* are also found in AML, leading to increased expression of *EVI1* comparable to that of 3q26-rearranged cases [167]. It has recently come to our attention that an enhancer region downstream of *MECOM* can become amplified in ovarian cancer, with similar effects (Stefan Gröschel, personal communication). It is thus possible that high expression of *EVI1* is driven by an analogous mechanism in AML. Alternatively, loss of silencers or insulators by small deletions not detected so far could achieve the same effect. It cannot be ruled out that a permissive chromatin environment enables the expression of *EVI1*, as is the case of HSPCs in normal hematopoiesis. In order to pinpoint the underlying mechanism, appropriate sequencing experiments should be conducted in a cohort of *EVI1*-positive AML cases without 3q26 rearrangements. Structural variants, point mutations and indels could be detected by 3q-capture or WGS, which should be complemented by epigenomics data (ChIP-seq, ATAC-seq) to assess their functional significance, followed by validation in cell lines. Cryptic gene fusions could also be investigated by RNA-seq. In the absence of any genomic variants, supervised comparisons of ChIP-seq, ATAC-seq or methylation (e.g. targeted bisulfite sequencing) between *EVI1*-positive and *EVI1*-negative cases could unveil differences at the epigenetic level in the vicinity of the *EVI1* locus.

2.3 Epigenetic dysregulation driving altered gene expression in AML

Transcriptional control involves a complex network of regulatory elements that interact with each other, including *cis*-regulatory elements (CREs) like promoters or enhancers, as well as *trans*-regulatory elements like TFs or chromatin modifiers. Alterations in components of this network are common in AML, but research traditionally focused on mutations in coding genes, including epigenetic regulators and transcription factors. In recent years, many examples of alterations involving CREs have emerged as prime drivers of gene dysregulation. Promoters can be silenced by hypermethylation [168] or targeted by mutations which can either introduce new TFBS [169,170] or inactivate them [171]. Aside from translocations leading to enhancer hijacking, discussed above, other enhancer-related mechanisms include the formation of novel enhancers by point mutations or indels [145], focal amplifications of existing enhancers [172] and disruption of architectural loops [173].

This thesis describes the involvement of several of these mechanisms in AML, often simultaneously. In **chapter 6**, we showed how allele-specific silencing of the *GATA2* promoter in AML with *CEBPA* DM is accompanied by hyperactivation of the -110 kb enhancer, presumably at the other allele. This is not only an example of two co-occurring epimutations affecting a single gene, but also of the intricate interactions between regulatory elements. In this case, aberrant overexpression driven by an enhancer can be compensated by silencing

of the promoter. In **chapter 7**, we showed that widespread hypermethylation in a subset of leukemias leads to silencing of key myeloid promoters and reshaping of genome architecture due to loss of CTCF binding.

2.3.1 Strategies to screen for oncogenic alterations in cis-regulatory regions

The identification of alterations in non-coding regions leading to cancer can be a daunting prospect. These regions account for roughly 98.5% of the genome [174,175], much of which is essentially junk DNA without function [176,177]. In fact, it is estimated that only around 5% of the genome is under purifying selection [178,179], meaning that mutations may randomly accumulate in a neutral or nearly neutral manner in the remaining 95% [180]. Therefore, it is not viable to conduct whole genome sequencing (WGS) and study every somatic variant in non-coding regions, as this would be tantamount to look for the proverbial "needle in a haystack". Besides, such an approach would be blind to epigenetic alterations that leave the DNA sequence intact. A second obstacle is to demonstrate that a mutation in a CRE effectively leads to aberrant expression of a gene under its control, which is not necessarily close in the linear genome.

We propose two possible approaches to tackle the aforementioned challenges, both of which have been used in this thesis: CRE-centric strategies and gene-centric strategies.

A CRE-centric strategy to identify cis-regulatory alterations

A conventional strategy to detect causal variations in regulatory regions associated with cancer is to focus the search on previously defined CREs, followed by additional studies to link those changes to gene dysregulation and functional studies. A proposed pipeline, partially adapted from [181], is presented here:

1. Restriction of search space: existing knowledge on regulatory regions can be used to narrow down the search for variants. The combination of ATAC-seq and H3K27ac ChIP-seq can reveal putative CREs active in a certain tissue, which can then be classified as promoters or enhancers on the basis of their proximity to TSSs or by integration with other histone marks like H3K4me3 (promoters) or H3K4me1 (enhancers). Other regulatory regions of importance are binding sites for structural proteins, such as CTCF.

2. Detection of somatic CRE mutations: changes in CREs can be genetic or epigenetic in nature, which calls for separate sets of approaches. The identification of one of such events may be interesting, but recurrence is a strong indicator of their somatic involvement. For example, insertions creating a super-enhancer upstream of *TAL1* occur in ~5% of T-ALL patients [145].

- **Genetic hits:** point mutations, copy number alterations or structural variants overlapping with CREs can be detected by WGS or capture DNA-seq. It is also possible to use ChIP-seq or ATAC-seq reads to identify genetic variation leading to increases in enhancer activity [146], but this strategy may fail for variants that completely abrogate TF binding.

- **Epigenetic hits:** changes in methylation at promoters, associated with gene silencing, can be detected by techniques such as MCIP-seq or whole genome bisulfite sequencing (WGBS). Gains or losses in promoter or enhancer activity can be measured by ATAC-seq or H3K27ac ChIP-seq.

3. Identification of the target gene (mutations in enhancers): only between 30% and 60% of enhancers interact with their closest promoters; the remaining enhancers regulate genes that are often located several hundreds of kb or even Mb away [121]. 3C-derived technologies like 4C or Hi-C are a valuable tool to establish the most likely targets of the dysregulated enhancer. In some instances, novel (i.e. somatic) interactions may be formed as a result of an oncogenic event, such as the translocation of an enhancer.

4. Detection of cancer-specific patterns of expression: mutations in a CRE are only likely to have a functional impact if they affect the expression of a coding gene. Transcriptomics data, such as RNA-seq, should be employed to assess whether the identified mutations alter the expression levels of a target gene. Furthermore, if heterozygous SNPs are present in the gene of interest, ASE should confirm that only the mutated allele is affected.

5. Testing in *in vitro* or *in vivo* models: the effects of mutations detected in patient samples or cell lines should be replicable in suitable models. CRISPR-Cas9, which has become an essential tool for genome editing, can be used to disrupt the function of a CRE or replace it with another [182]. Deactivated Cas9 (dCas9) can silence a CRE without introducing any genetic mutations, an approach known as CRISPRi [183]. Modifications of this procedure strengthen this effect by coupling dCas9 to chromatin modifier domains such as KRAB [184] and DNMT3A [185].

Variations of this proposed pipeline have been applied throughout the present thesis. In **chapter 3**, we performed 3q-capture to detect somatic SVs involving super-enhancers in regions translocated to *EVI1*, some of which were recurrent. Increased expression of *EVI1* was measured by RNA-seq. In **chapter 4**, we demonstrated that the *MYC* SE is a causal driver of *EVI1* overexpression by 1) introducing the t(3;8) into K562 cells, and 2) deleting critical components of the *MYC* SE. In **chapter 7**, we detected changes in methylation in promoters and CTCF binding sites, which we related to changes in gene expression and the in 3D genome structure with RNA-seq and Hi-C respectively.

A gene-centric survey of cis-regulatory alterations

The strategy employed in **chapter 6** was gene-centric, rather than CRE-centric. That is, instead of screening for changes in CREs and relating them to genes under their control, **we surveyed the transcriptome for signs of deregulated gene expression**. To this end, we relied on ASE as a surrogate marker for changes in CREs, under the assumption that those typically affect a single allele (Figure 4). Previous studies have confirmed that ASE can be a powerful indicator of CRE dysregulation; for example, in t(3;3)/inv(3) AML both *GATA2* and *EVI1* both exhibit monoallelic expression [4]. Indeed, in a cohort of ~200 AMLs we detected ASE of genes involved in known chromosomal rearrangements, such as inv(16) or t(8;21), which confirmed the validity of our approach. More importantly, we found recurrent ASE

of various cancer-related genes, most prominently *GATA2*. Even though *GATA2* ASE was present in almost 60% of tested AMLs, there was a strong association with *CEBPA* DM.

The next step in this strategy is to trace the ASE signal back to the corresponding CREs, which can be achieved with chromatin interaction data, as suggested above, or prior knowledge of gene regulation. In this case, we focused on the *GATA2* promoter and the -110 kb enhancer to establish that ASE is the result of promoter hypermethylation on one allele and enhancer hyperactivation on the other allele in *CEBPA* DM AML. As a result, *GATA2* transcript levels remain unchanged with respect to other AMLs. Long-read sequencing with Nanopore was critical to demonstrate that *GATA2* promoter hypermethylation and loss of expression take place on the same allele.

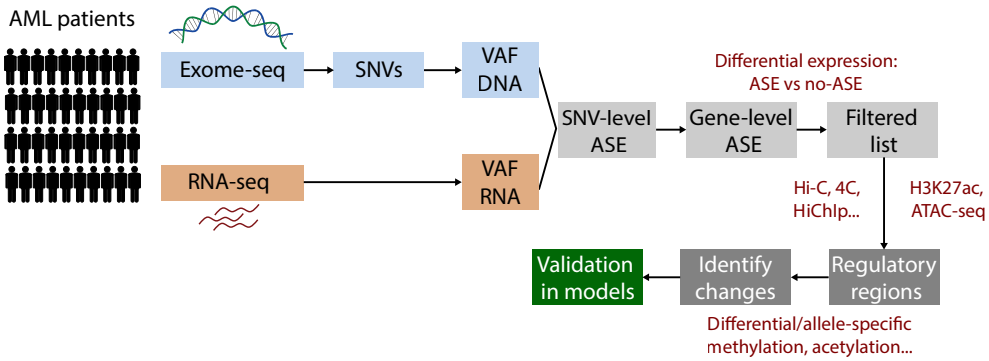


Figure 4. Detection of allele-specific expression in AML, indicative of alterations in CREs.

This approach allows mechanism-agnostic detection of potentially deregulated genes, which otherwise may be challenging and require the integration of multiple datasets. An ASE screen can be performed with only RNA-seq, albeit combination with WES or WGS is highly advisable to eliminate confounding factors such as CNAs or subclonal mutations, and to discriminate monoallelic expression from homozygous variants. Nevertheless, it is limited to regions that harbor at least one heterozygous SNV and it may also be challenging to pinpoint the causal mechanism behind ASE. Besides, it requires very careful filtering of genetic variants, since technical artifacts in certain regions with poor mappability can introduce false positives. For this reason, this exercise is often conducted with SNVs only, since other variants like indels are more prone to alignment errors.

2.3.2 A central role for GATA2 in acute myeloid leukemia

A recurrent observation in several chapters of this thesis is the involvement of GATA2 in multiple pathogenic processes. Aside from the loss of the *GATA2* enhancer in t(3;3)/inv(3) AML, we also detected frequent CNLs of the *GATA2* gene in atypical 3q26, as well as unexplained ASE (**chapter 3**). As discussed in section **2.2.2**, this points to a likely participation

of GATA2 defects in 3q26-rearranged AML. Surprisingly, *GATA2* ASE was present in more than 60% AML cases, many of which did not harbor either 3q26 aberrations or *CEBPA* DM (**chapter 6**). Moreover, GATA2 is one of the heptad of hematopoietic TFs that bind super-enhancers commonly translocated to *EVI1* (**chapter 3**), including the *GATA2* super-enhancer itself. In fact, the p300 region critical for *EVI1* expression in inv(3)/t(3;3) contains motifs for TFs of the GATA family (**chapter 5**). All in all, GATA2 emerges as a pivotal contributor to leukemogenesis in AML. But what makes GATA2 so important?

Functional significance of GATA2 in hematopoiesis

GATA2 is a member of the GATA family of zinc-finger (ZF) TFs, whose name derives from their ability to bind the (A/T)GATA(A/G) consensus sequence, also denoted as WGATAR [186]. The founding member of the family was GATA1, also referred to as NF-E1 or Eryf1, independently discovered as a key activator of globin gene expression in erythroid cells by multiple groups in 1988 [186–190]. The cloning of GATA1 revealed that interaction with this motif is mediated by a C-terminal ZF (ZF2) [191,192], whereas another N-terminal ZF (ZF1) increases binding stability and facilitates interaction with other proteins such as Friend of GATA (FOG) [193]. Soon afterwards, GATA2 and GATA3 were identified by homology with GATA1, first in chicken [194] and later in humans [195,196]. Interplay between these different GATA proteins takes place in the form of a "GATA switch" whereby one of them replaces another at key stages of differentiation [197]. For example, GATA1 displaces GATA2 from the *GATA2* promoter in erythropoiesis, leading to transcriptional repression [198].

GATA2 is indispensable for HSC proliferation and survival [130,199], as well as for HSC generation in the embryo [131,200] and GMP function [139]. Accordingly, expression of *GATA2* can be detected in HSCs, early myeloid progenitors and erythroid cells [201]. Transcription of *GATA2* is controlled by a number of enhancers that also act as "GATA switch sites", including an intronic +9.9 kb enhancer (+9.5 in mice), several proximal enhancers and a distal -110 kb (-77 in mice) enhancer [202]. While the proximal enhancers are dispensable for *Gata2* expression and hematopoiesis [203], loss of either the +9.5 kb [204] or the -77 kb [205] enhancers dramatically reduces *Gata2* levels and disturbs hematopoiesis. In particular, the -77 kb element is mainly involved in *Gata2* expression in myeloid commitment, whereas the +9.5 kb enhancer regulates HSC emergence [202].

In keeping with its essential role in hematopoiesis, **a fine control of *GATA2* expression levels must be maintained to ensure proper hematopoietic function.** In *Gata2*$^{+/-}$ mouse models, low *Gata2* expression compromises HSC homeostasis [138] and GMP function [139]. In humans, haploinsufficiency resulting from inactivating mutations in *GATA2* coding regions or its +9.9 kb enhancer causes inheritable disorders like the MonoMAC and the Emberger syndromes [133–136], currently considered manifestations of a single entity known as "*GATA2* deficiency" [206]. Patients with these defects exhibit various cytopenias and frequent infections and are at risk for developing familial MDS and AML [137]. Complete loss of *Gata2* is deleterious

not only for HSCs, but also for LSCs, which undergo apoptosis as a result of decreased Bcl-2 levels [132]. On the other hand, overexpression of *GATA2* blocks proliferation and differentiation of HSCs and progenitors, fostering quiescence [207,208]. At lower levels, enforced expression of *GATA2* inhibits lymphoid development at the CLP stage, but enhances GMP self-renewal and myelopoiesis [209]. Comparably high levels of *GATA2* have been linked to poor prognosis in AML [210,211]. In sum, perturbations in the physiological levels of *GATA2*, either by excess or defect, disrupt hematopoiesis.

Alterations in GATA2 are key drivers of leukemogenesis

Somatic mutations of *GATA2* are found in ~3% of sporadic AMLs [212,213] and are clustered in the two ZF domains [214]. Lesions in the ZF1 region are more frequent in AML and are associated with *CEBPA* DM [6] and inv(3)/t(3;3) [91], as well as better prognosis [215]. In contrast, the C-terminal ZF2 is mainly targeted by somatic mutations in CML with blast crisis [216] or by germline alterations in familial AML [137]. Although it was traditionally believed that ZF2 mutations abolish *GATA2* DNA binding activity, resulting in haploinsufficiency, recent studies depict a more nuanced picture, with both loss-of-function and gain-of-function outcomes [217,218]. Transcriptional activation and sometimes DNA binding are impaired by certain ZF1 mutations (like L321F), but the mechanisms remain unclear.

The high frequency of *GATA2* mutations and *GATA2* ASE in AML with *CEBPA* DM points to an **interplay between *CEBPA* and *GATA2* in leukemogenesis**. Hinting at a possible mechanism, Greif and colleagues reported that GATA2 ZF1 mutant proteins exhibit reduced cooperation with CEBPA, leading to downregulation of CEBPA targets [6]. The fact that mutant *GATA2* alleles are preferentially expressed in *CEBPA* DM AML, at the expense of their wild type counterparts, probably exacerbates this phenomenon (**chapter 6**). In healthy hematopoiesis, the sequence in which in *GATA2* and *CEBPA* are expressed controls fate choice among branches of myelopoiesis [219], further suggesting an interaction between these two TFs.

It is unclear **how *GATA2* ASE participates in AML development in the absence of *GATA2* mutations**, but we speculate that it is related to altered transcript levels in previous stages. Moderate upregulation of *GATA2* by a hyperactive enhancer may create a fertile ground for leukemogenesis by accelerating progenitor self-renewal [209], but loss of expression by compensatory methylation of the other allele may be favorable in later stages (Figure 5). This hypothesis is in line with findings in mice with inv(16), which exhibit upregulated *Gata2* in preleukemic cells, but acquire *Gata2* deletions in a leukemic phase [220]. Interestingly, *GATA2* mutations and haploinsufficiency also co-occur with *CEBPA* downregulation driven by *EVI1* overexpression in 3q26-rearranged AML. An intriguing possibility is that altered GATA2 levels synergize with CEBPA dysfunction by disrupting their cooperation at sensitive promoters, as is the case for GATA2 mutations. However, we did not detect an association between *GATA2* ASE and other subtypes with reduced *CEBPA* expression, such as t(8;21) or

single *CEBPA* mutations. It may prove challenging to reproduce this mechanism in a model, but it would be interesting to examine whether mice with *Cebpa* DM [221] also acquire *GATA2* ASE and, if so, whether it accelerates leukemogenesis.

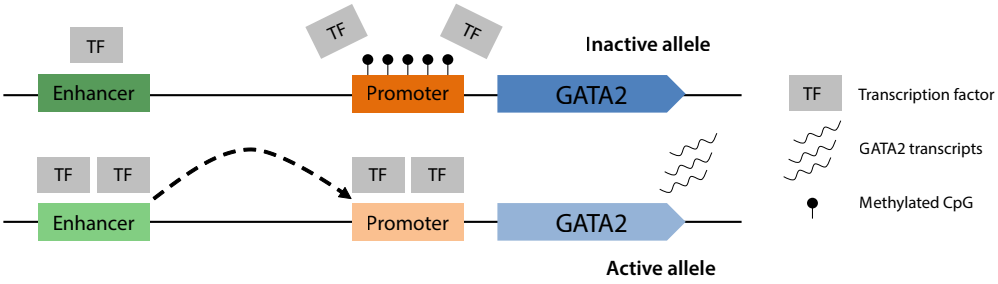


Figure 5. Mechanisms driving GATA2 ASE in *CEBPA* DM AML.

The exact nature of the interplay between *GATA2* and *CEBPA* is only one in many unanswered questions. **What is the order of acquisition of *GATA2* ASE and *CEBPA* mutations**? While the small VAF of *GATA2* mutations relative to *CEBPA* DM unequivocally defines them as subclonal, such a relationship cannot be established for *GATA2* ASE because gene expression is not a dichotomous variable. That is, for a given transcript frequency of allele A (e.g. 20%), we cannot distinguish a situation in which all cells exhibit reduced levels of that allele (20:80 A/B ratio) from a complete silencing in only a fraction of the cells (only 4 in 10 cells express A). However, the fact that *GATA2* ASE is present in almost all the *CEBPA* DM cases and in a large fraction of other AMLs strongly suggests that *GATA2* ASE precedes at least the acquisition of *GATA2* mutations, which are favored in the expressed allele (Figure 6). We further hypothesize that *GATA2* ASE is in fact a preuleukemic event that spontaneously takes root in HSCs or progenitor cells, fostering the development of AML in cooperation with mutations like *CEBPA*. This hypothesis should be tested using full-length single cell transcriptomics in AML with *CEBPA* DM, enabling the simultaneous tracking of both *CEBPA* mutations and *GATA2* ASE in the same patient. Besides, mice with *CEBPA* DM could be used to confirm whether *GATA2* mutations indeed follow *GATA2* ASE [221]. Interestingly, we detected *GATA2* ASE in both remission and diagnosis samples of a single *CEBPA* DM patient, but we could not establish if it was already present in the germ line. On the other hand, familial *CEBPA* mutations [222] necessarily precede *GATA2* ASE, but it can be argued that single *CEBPA* mutations are not leukemogenic. In AML with inv(3)/t(3;3), both *GATA2* haploinsufficiency and *CEBPA* downregulation are acquired simultaneously.

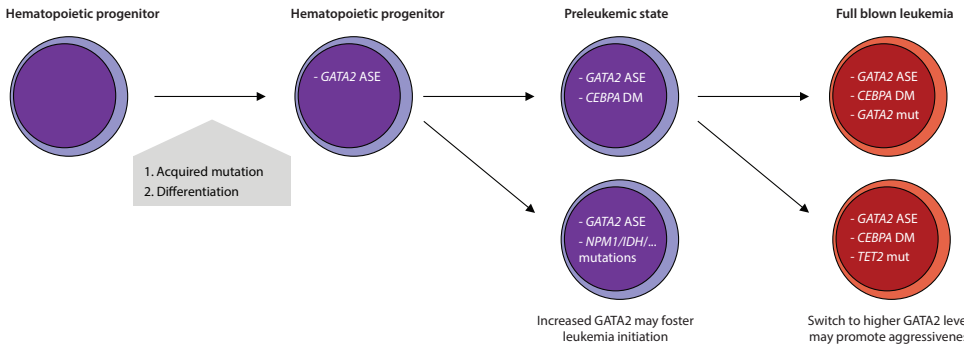


Figure 6. Proposed order of acquisition of *GATA2* ASE in *CEBPA* DM AML.

Other studies have also reported hypermethylation of the *GATA2* promoter in normal karyotype AML, in the absence of any *CEBPA* mutations [223,224]. What is the role of *GATA2* ASE in other AMLs without *CEBPA* alterations? In the same way GATA2 and CEBPA cooperate to regulate gene expression, other hematopoietic TFs are similarly dependent on GATA2. These include SCL, LYL1, LMO2, RUNX1, ERG, and FLI-1, the components of a "heptad" that colocalize with GATA2 and engage in protein-protein interactions with it [119]. The cooperation between RUNX1 and GATA2 was proven by the synthetic lethality of Gata2+/– and Runx1+/– in mice. Although the association was not significant, we also observed *GATA2* ASE in ~70% of AMLs with RUNX1. Besides, as a pivotal hematopoietic regulator, GATA2 controls the expression of many genes frequently mutated or dysregulated in AML. For example, the expression of the tumor suppressor *WT1*, mutated in around ~8% of AMLs [212,213], is controlled by binding of GATA2 or GATA1 to a 3' enhancer. We detected *GATA2* ASE in 90% of the patients with *WT1* mutations, which are a sign poor prognosis [225]. To summarize, *GATA2* ASE may cooperate with other lesions by either disrupting protein-protein interactions or modulating transcription of the affected genes. Validation of this hypothesis may be conducted by measuring synergistic effects between mutations in cancer-associated genes and either haploinsufficiency or forced overexpression of *GATA2* in cell lines or mouse models.

2.3.3 The road not taken: other candidates from the ASE screen

The screen performed in **chapter 6** identified *GATA2* as the most recurrent gene with ASE among those known to be mutated in cancer (as per the COSMIC database [226]) or involved in myelopoiesis. However, several other interesting candidates for further research were **also present in a large number of samples**, such as *THBS1* (29%) or *CDKN2A* (11%). The gene product of *THBS1*, thrombospondin-1, is a glycoprotein that mediates cell adhesion, with multiple roles in cancer such as invasion, migration, proliferation, apoptosis and tumor immunity [227]. The THBS1 receptor CD47 confers protection against macrophage-mediated

clearance (a “don’t eat me” signal) and is highly expressed on circulating HSCs and LSCs, which co-opt this mechanism to evade phagocytosis [228]. High CD47 expression on LSCs is an adverse prognostic factor in AML [229], and blockade of CD47 by antibodies like magrolimab has shown promise in clinical trials [230]. A study in 116 AML patients determined that low levels of THBS1, possibly caused by promoter hypermethylation, correlated with lower survival [231]. In our cohort, cases with *THBS1* ASE exhibited higher *THBS1* expression than other AMLs (1.7 fold-change), though the difference was not statistically significant (adjusted p-value = 0.26). This seemingly rules out a silencing mechanism, unless it is compensated on the other allele as reported for *GATA2*. The levels of *THBS1* were higher than in CD34+ cells from healthy donors regardless of whether ASE was present, also excluding the possibility that cases without ASE exhibited silencing on both alleles. Moreover, *THBS1* ASE was strongly associated with inv(16), as it was present in 9 out of 11 cases (p-value = 0.0008). The *CDKN2A* gene encodes two tumor suppressor proteins that regulate the cell cycle: $p16^{INK4A}$ and $p14^{ARF}$ [232]. This gene is frequently repressed by epigenetic mechanisms in AML, resulting in accelerated proliferation, including the deposition of H3K27me3 marks by the polycomb group [11,233,234] and DNA hypermethylation in AML with IDH1 mutations, [235]. Low expression of C*DKN2A* is an independent predictor of poor survival in AML patients of advanced age [236]. However, patients with *CDKN2A* ASE in our study did not exhibit reduced expression of the gene, suggesting a different mechanism.

Other candidates from the screen were less frequent, but involved **critical regulators of lineage commitment**, like *IRF8* (4%) or *MEIS1* (4%). Interferon regulatory factor 8 (IRF8) favors the production of monocytes and dendritic cells at the expense of neutrophils [237] and its inactivation in mice models leads to a leukemia-like syndrome [238]. Upregulation of *IRF8* has been detected in various subsets of AML and is a sign of poor prognosis [239,240], possibly because it supports aberrant proliferation [241]. MEIS1, a cofactor of HOX proteins, preserves HSC quiescence [242,243] and is upregulated in AML with NPM1 mutations [244] and with MLL rearrangements [245], mostly co-occurring with *HOXA9* overexpression [246]. The collaboration of MEIS1 with HOXA9 [247,248] or with *NPM1* mutations [249] is sufficient to induce leukemia in mice. Interestingly, C/EBPA is a critical collaborator in leukemogenesis induced by HOXA9 and MEIS1 [250].

Moreover, **genes without known mutations linked to cancer, but also with highly recurrent ASE, may be of interest**. Some examples include *IRF5* (43%), a target of p53 for induction of apoptosis in HSCs [251], and *CD13/ANPEP* (21%), a myeloid surface marker typically expressed on most AML blasts [252]. Of note, although *ANPEP* was not previously considered because mutations in this gene have not been reported in cancer, its relationship with AML has been documented. Thus, it may be preferable to use the DisGeNET database [253] for candidate selection, as it contains genes associated with AML mined from the scientific literature. On the other hand, genes without previous evidence of involvement in leukemia also offer the greatest potential for novel discoveries. To ensure that the allelic bias in transcription can be related to gains or losses in nearby CREs, differential expression

between AML cases with and without ASE can be used to prioritize these candidates (Table 2). For example, *MECOM* is overexpressed in patients with *MECOM* ASE because those often harbor 3q26 rearrangements leading to upregulation of the *EVI1* isoform encoded by this gene.

Identification of abnormalities in the regulation of any of these proposed candidates can be conducted in a systematic manner by integration of genomics and epigenomics data. We have generated H3K27ac ChIP-seq and ATAC-seq data for the same AML patients that were sequenced for the ASE screen, which can reveal gained or loss enhancers in the vicinity of these genes. Nanopore-based sequencing can be employed to simultaneously measure methylation at promoters and relate the signal to the expressed allele, taking advantage of long read technology [254].

Table 2. List of genes with ASE in AML associated with either gain (fold change > 0) or loss (fold change < 0) of expression.

Gene name	Cancer	Hematopoiesis	Leukemia	Fraction of cases	log2 fold change	Adjusted p-value
PARVB	NO	NO	NO	4%	-2.19	0.00
DAPK1	NO	NO	YES	5%	-2.00	0.00
CARD9	NO	NO	NO	6%	-1.41	0.00
CLEC11A	NO	NO	NO	31%	-1.25	0.05
PXDN	NO	NO	YES	9%	-1.16	0.02
USP6	YES	NO	NO	5%	-1.08	0.00
MYH11	YES	NO	YES	10%	2.73	0.02
EPCAM	NO	NO	YES	11%	2.91	0.00
AFDN	NO	NO	YES	4%	3.33	0.00
SYCP2L	NO	NO	NO	7%	3.35	0.04
FOXC1	NO	YES	YES	24%	3.47	0.00
MEGF10	NO	NO	NO	4%	3.51	0.00
FN1	NO	NO	YES	8%	3.87	0.00
CACNA1H	NO	NO	NO	4%	3.93	0.00
MECOM	YES	YES	YES	7%	4.04	0.00
TACSTD2	NO	NO	NO	5%	4.22	0.00
C2CD4B	NO	NO	NO	5%	4.68	0.00
DEFB1	NO	NO	YES	13%	5.94	0.00

Only a selection of candidates with significant changes in expression (p-value < 0.05) is shown here. The columns Cancer, Hematopoiesis and Leukemia indicate if these genes are involved in these processes according to COSMIC, GO and DisGeNet respectively. The fold change and p-value were calculated with DESeq2.

2.3.4 CEBPA is a crucial determinant of cell identity in benign and malignant hematopoiesis

Lymphoid differentiation is associated with higher levels of DNA methylation at the promoters and binding sites of myeloid TFs [255,256]. Furthermore, mouse models with *Dnmt1* [257] or *Dnmt3a* [258] deficiencies exhibit myeloid skewing. Similarly, polycomb-mediated repression seems to be dispensable for myelopoesis, as mice without *Ezh1* [259] or *Ezh2* [260] only exhibit compromised lymphocyte production. Altogether, the evidence suggests that the myeloid program is active by default unless it is actively repressed. ***CEBPA* is one of the critical myeloid factors that must be repressed as a part of this cell fate decision**, as overexpression of *CEBPA* is sufficient to enforce a myeloid program in lymphoid progenitors [219] and reprogram B-cells into macrophages [261]. On the other hand, absence of *CEBPA*

results in failure of myelopoiesis at the transition from CMP to GMP [11]. Indeed, *CEBPA* is only expressed in myeloid progenitors and terminally differentiated cells, but not in their lymphoid counterparts [28,29].

Although murine *Cebpa* is differentially expressed between lymphoid and myeloid progenitors, this phenomenon is not accompanied by promoter methylation, pointing at other silencing mechanisms [255]. We confirmed similar observations in human terminally differentiated cells using data from the Blueprint consortium [262], albeit a few CpGs were uniquely methylated in a few lymphoid cells (Figure 7). We did not observe consistent differences in H3K27me3 levels between lymphoid and myeloid cells either, excluding a possible involvement of the polycomb group in *CEBPA* repression (Figure 8). Instead, studies in mice [263] and in humans [2] have revealed that *CEBPA* regulation along differentiation is controlled by the +42 kb enhancer, which contains binding sites for hematopoietic TFs (see also **section 2.1**). This region is active in HSCs, early multipotent progenitors and myeloid cells, but not in terminally differentiated lymphoid cells [263]. Indeed, H3K27ac ChIP-seq data from Blueprint only showed activity of *CEBPA* enhancers in myeloid cells (Figure 9). Thus, **expression of *CEBPA* seems to be regulated by its enhancers rather than by silencing mechanisms in normal hematopoiesis**.

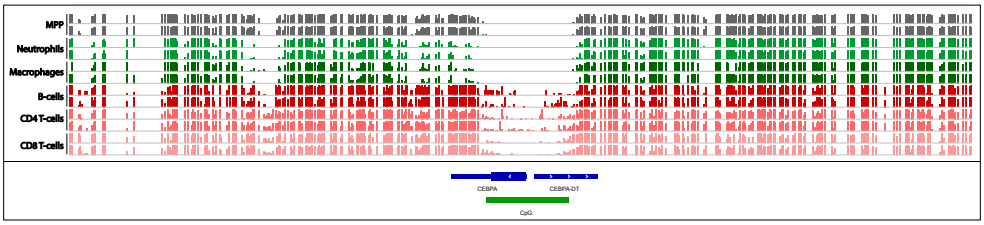


Figure 7. Methylation of the *CEBPA* promoter in different hematopoietic cell types.

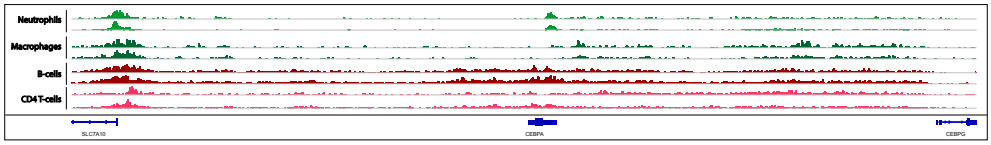


Figure 8. H3K27me3 levels at *CEBPA* and neighboring regions in different hematopoietic cell types.

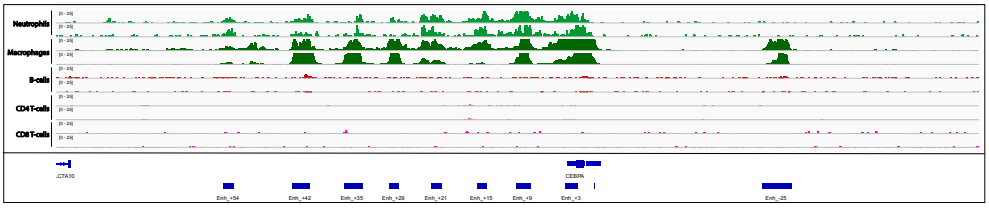


Figure 9. H3K27ac levels of *CEBPA* enhancers in hematopoietic cell types. Enhancer elements are indicated below in blue.

The group of leukemias referred to as CIMP exhibit hypermethylation-driven repression of several TFs involved in lineage commitment, including *CEBPA*, and mixed T-lymphoid/myeloid phenotype (**chapter 7**). In contrast with normal tissues, the ***CEBPA* promoter is aberrantly methylated in T-ALL and CIMP leukemias** (Figure 10). As a result, *CEBPA* is locked in an inactive state that completely abrogates myelopoiesis, even in the presence of enhancer activation. CIMP leukemias were originally identified because they clustered with *CEBPA* DM AML in unsupervised analyses of microarray expression data, yet had no detectable *CEBPA* levels, leading to the observation that the *CEBPA* promoter was methylated [8]. Likewise, we observed that CIMPs preferentially clustered with *CEBPA* DM AML when analyzing RNA-seq, ATAC-seq or H3K27ac ChIP-seq datasets from AML and T-ALL patients (**chapter 7**). Both groups were located equidistantly from T-ALL and the rest of AMLs in a two-dimensional visualization of these data. CIMPs also exhibited similarities with t(8;21) AML, in which *CEBPA* is also repressed by this oncoprotein [80]. Altogether, these results imply that **silencing or loss of *CEBPA* function is a major determinant of the epigenetic landscape in leukemia**. In line with this hypothesis, blasts with *CEBPA* DM often express T-cell genes such as *CD7* and *TRD* [264].

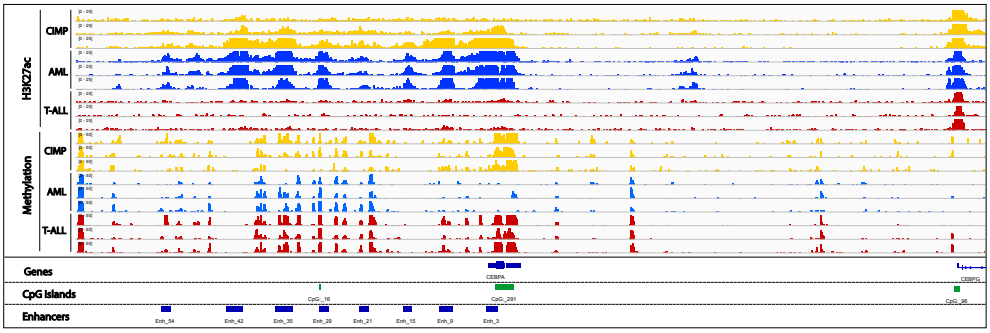


Figure 10. Methylation and acetylation of the *CEBPA* locus in various leukemia groups.

How is the methylation of *CEBPA* established in leukemia? The methylation pattern of CIMP leukemias is clearly distinct from that of myeloid leukemias, including most *CEBPA* DM cases, yet very similar to that of T-ALL. This seems to indicate that loss of *CEBPA* is a secondary event related to lymphoid development. However, hypermethylation of *CEBPA* is exclusive of leukemia, whereas expression in normal tissues is controlled by other mechanisms. In fact, global methylation levels in leukemic cells are higher than in any of their healthy counterparts. This strongly suggests that aberrant methylation of *CEBPA* is a cancer-specific process that occurs both in T-ALL and CIMP leukemias. As it is independent from mutations in the methylation machinery, this process may require transient dysregulation in these enzymes. Microwaves of methylation and demethylation take place during differentiation [265], so **a transient dysfunction in the control of the methylation machinery could explain this phenotype**. The downregulation of *TET2* coupled with upregulation of *DNMT3A/B*

in CIMP leukemias supports this hypothesis (**chapter 7**), but other mechanisms may be involved, as AMLs with *TET2* mutations do not present this extreme hypermethylation.

Lack of *CEBPA* itself may also contribute this methylator phenotype, given the pivotal role of this TF in shaping the epigenome of CIMP leukemias. Although the methylation pattern in *CEBPA* DM AML is different than in CIMP leukemias, this discrepancy could be explained by a residual function of *CEBPA* with N-terminal mutations, which results in the transcription of a shorter p30 isoform that lacks a transactivation domain, but is capable of DNA binding and can therefore exert other functions [266]. In fact, we detected a single AML case with double C-terminal mutations, leading to a complete loss of DNA binding, that clustered together with CIMP and T-ALL cases in the analysis of methylation data. Furthermore, some AMLs with *CEBPA* DM are relatively hypermethylated, though not to the same extent as the CIMP cases [267,268]. It will be necessary to confirm this hypothesis by carrying out methylation assays in additional AML cases with double C-terminal mutations.

How could *CEBPA* dysfunction lead to hypermethylation? We hypothesize that loss of *CEBPA* in early progenitors introduces a differentiation block that precludes myelopoiesis, forcing cells to adopt a lymphoid program. Disruptions in the methylation machinery, like repression of *TET2*, may exacerbate the waves of *de novo* methylation that accompany this fate choice, eventually resulting in aberrant genome-wide methylation. The details of this process remain highly speculative. Single cell methylation studies in patients could reveal the order of acquisition of methylation patterns in leukemic and/or preleukemic clones. Although material from CIMP leukemias is scarce, it should be possible to study how T-ALL also acquires this aberrant patterns that markedly differ from healthy terminally differentiated cells. Experiments in *in vitro* or *in vivo* models should be conducted to evaluate whether loss of *CEBPA* is indeed sufficient to reshape the epigenome and whether any clones with altered methylation develop. An important challenge in these studies might be the intimate relationship between *CEBPA* loss and differentiation, as the effects proposed above may only become apparent in a specific cell of origin.

REFERENCES

1. Hasemann, M. S. *et al.* C/EBPα Is Required for Long-Term Self-Renewal and Lineage Priming of Hematopoietic Stem Cells and for the Maintenance of Epigenetic Configurations in Multipotent Progenitors. *PLoS Genet.* **10**, (2014).
2. Avellino, R. *et al.* An autonomous CEBPA enhancer specific for myeloid-lineage priming and neutrophilic differentiation. *Blood* **127**, 2991–3003 (2016).
3. Guo, H., Cooper, S. & Friedman, A. D. In vivo deletion of the Cebpa +37 kb enhancer markedly reduces Cebpa mRNA in myeloid progenitors but not in non- hematopoietic tissues to impair granulopoiesis. *PLoS One* **11**, 1–23 (2016).
4. Gröschel, S. *et al.* A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in Leukemia. *Cell* **157**, 369–381 (2014).
5. Yamazaki, H. *et al.* A remote GATA2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating EVI1 expression. *Cancer Cell* **25**, 415–427 (2014).
6. Greif, P. A. *et al.* GATA2 zinc finger 1 mutations associated with biallelic CEBPA mutations define a unique genetic entity of acute myeloid leukemia. *Blood* **120**, 395–403 (2012).
7. Figueroa, M. E. *et al.* Genome-wide epigenetic analysis delineates a biologically distinct immature acute leukemia with myeloid/T-lymphoid features. *Blood* **113**, 2795–2804 (2009).
8. Wouters, B. J. *et al.* Distinct gene expression profiles of acute myeloid/T-lymphoid leukemia with silenced CEBPA and mutations in NOTCH1. *Blood* **110**, 3706–3714 (2007).
9. Gebhard, C. *et al.* Profiling of aberrant DNA methylation in acute myeloid leukemia reveals subclasses of CG-rich regions with epigenetic or genetic association. *Leukemia* **33**, 26–36 (2019).
10. Heath, V. *et al.* C/EBPα deficiency results in hyperproliferation of hematopoietic progenitor cells and disrupts macrophage development in vitro and in vivo. *Blood* **104**, 1639–1647 (2004).
11. Zhang, P. *et al.* Enhancement of hematopoietic stem cell repopulating capacity and self-renewal in the absence of the transcription factor C/EBPα. *Immunity* **21**, 853–863 (2004).
12. Zhang, D. E. *et al.* Absence of granulocyte colony-stimulating factor signaling and neutrophil development in CCAAT enhancer binding protein α-deficient mice. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 569–574 (1997).
13. Radomska, H. S. *et al.* CCAAT/enhancer binding protein alpha is a regulatory switch sufficient for induction of granulocytic development from bipotential myeloid progenitors. *Mol. Cell. Biol.* **18**, 4301–4314 (1998).
14. Wang, D., D'Costa, J., Civin, C. I. & Friedman, A. D. C/EBPalpha directs monocytic commitment of primary myeloid progenitors. *Blood* **108**, 1223–1229 (2006).
15. Ma, O., Hong, S. H., Guo, H., Ghiaur, G. & Friedman, A. D. Granulopoiesis requires increased C/EBPα compared to monopoiesis, correlated with elevated Cebpa in immature G-CSF receptor versus M-CSF receptor expressing cells. *PLoS One* **9**, (2014).
16. Wang, D., D'Costa, J., Civin, C. I. & Friedman, A. D. C/EBPα directs monocytic commitment of primary myeloid progenitors. *Blood* **108**, 1223–1229 (2006).
17. Hendricks-Taylor, L. R. & Darlington, G. J. Inhibition of cell proliferation by C/EBP alpha occurs in many cell types, does not require the presence of p53 or Rb, and is not affected by large T-antigen. *Nucleic Acids Res.* **23**, 4726–33 (1995).
18. Johansen, L. M. *et al.* c-Myc Is a Critical Target for C/EBPα in Granulopoiesis. *Mol. Cell. Biol.* **21**, 3789–3806 (2001).
19. Wang, H. *et al.* C/EBPα arrests cell proliferation through direct inhibition of Cdk2 and Cdk4. *Mol. Cell* **8**, 817–828 (2001).

20. Tsukada, J., Yoshida, Y., Kominato, Y. & Auron, P. E. The CCAAT/enhancer (C/EBP) family of basic-leucine zipper (bZIP) transcription factors is a multifaceted highly-regulated system for gene regulation. *Cytokine* vol. 54 6–19 (2011).
21. Zhang, D. E. *et al.* CCAAT enhancer-binding protein (C/EBP) and AML1 (CBF alpha2) synergistically activate the macrophage colony-stimulating factor receptor promoter. *Mol. Cell. Biol.* **16**, 1231–1240 (1996).
22. Hohaus, S. *et al.* PU.1 (Spi-1) and C/EBP alpha regulate expression of the granulocyte-macrophage colony-stimulating factor receptor alpha gene. *Mol. Cell. Biol.* **15**, 5830–5845 (1995).
23. Smith, L. T., Hohaus, S., Gonzalez, D. A., Dziennis, S. E. & Tenen, D. G. PU.1 (Spi-1) and C/EBPα regulate the granulocyte colony-stimulating factor receptor promoter in myeloid cells. *Blood* **88**, 1234–1247 (1996).
24. Guo, H., Ma, O. & Friedman, A. D. The Cebpa +37-kb enhancer directs transgene expression to myeloid progenitors and to long-term hematopoietic stem cells. *J. Leukoc. Biol.* **96**, 419–426 (2014).
25. Ye, M. *et al.* C/EBPa controls acquisition and maintenance of adult haematopoietic stem cell quiescence. *Nat. Cell Biol.* **15**, 385–394 (2013).
26. Porse, B. T. *et al.* Loss of C/EBPα cell cycle control increases myeloid progenitor proliferation and transforms the neutrophil granulocyte lineage. *J. Exp. Med.* **202**, 85–96 (2005).
27. Traver, D. *et al.* Fetal liver myelopoiesis occurs through distinct, prospectively isolatable progenitor subsets. *Blood* **98**, 627–635 (2001).
28. Miyamoto, T. *et al.* Myeloid or lymphoid promiscuity as a critical step in hematopoietic lineage commitment. *Dev. Cell* **3**, 137–147 (2002).
29. Akashi, K., Traver, D., Miyamoto, T. & Weissman, I. L. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature* **404**, 193–197 (2000).
30. Chambers, S. M. *et al.* Hematopoietic Fingerprints: An Expression Database of Stem Cells and Their Progeny. *Cell Stem Cell* **1**, 578–591 (2007).
31. Månsson, R. *et al.* Molecular Evidence for Hierarchical Transcriptional Lineage Priming in Fetal and Adult Stem Cells and Multipotent Progenitors. *Immunity* **26**, 407–419 (2007).
32. Laurenti, E. *et al.* The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment. *Nat. Immunol.* **14**, 756–763 (2013).
33. Wölfler, A. *et al.* Lineage-instructive function of C/EBPα in multipotent hematopoietic cells and early thymic progenitors. *Blood* **116**, 4116–4125 (2010).
34. Spangrude, G. J., Heimfeld, S. & Weissman, I. L. Purification and characterization of mouse hematopoietic stem cells. *Science (80-.).* **241**, 58–62 (1988).
35. Morrison, S. J. & Weissman, I. L. The long-term repopulating subset of hematopoietic stem cells is deterministic and isolatable by phenotype. *Immunity* **1**, 661–673 (1994).
36. Orkin, S. H. & Zon, L. I. Hematopoiesis: An Evolving Paradigm for Stem Cell Biology. *Cell* vol. 132 631–644 (2008).
37. Clark, M. B. *et al.* The Reality of Pervasive Transcription. *PLOS Biol.* **9**, e1000625 (2011).
38. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
39. Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun. 2018 91* **9**, 1–9 (2018).
40. Wang, X., He, Y., Zhang, Q., Ren, X. & Zhang, Z. Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2. *Genomics. Proteomics Bioinformatics* (2021) doi:10.1016/j.gpb.2020.02.005.

41. Martin, E. W. *et al.* Chromatin accessibility maps provide evidence of multilineage gene priming in hematopoietic stem cells. *Epigenetics and Chromatin* **14**, 1–15 (2021).
42. Ranzoni, A. M. *et al.* Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis. *Cell Stem Cell* **28**, 472-487.e7 (2021).
43. Muench, D. E. *et al.* Mouse models of neutropenia reveal progenitor-stage-specific defects. *Nature* (2020) doi:10.1038/s41586-020-2227-7.
44. Zambetti, N. A. *et al.* Deficiency of the ribosome biogenesis gene Sbds in hematopoietic stem and progenitor cells causes neutropenia in mice by attenuating lineage progression in myelocytes. *Haematologica* **100**, 1285–1293 (2015).
45. Ding, L. & Morrison, S. J. Haematopoietic stem cells and early lymphoid progenitors occupy distinct bone marrow niches. *Nature* **495**, 231–235 (2013).
46. Zhang, J. *et al.* In situ mapping identifies distinct vascular niches for myelopoiesis. *Nature* **590**, 457–462 (2021).
47. Wei, Q. & Frenette, P. S. Niches for Hematopoietic Stem Cells and Their Progeny. *Immunity* vol. 48 632–648 (2018).
48. Lo Celso, C. *et al.* Live-animal tracking of individual haematopoietic stem/progenitor cells in their niche. *Nature* **457**, 92–96 (2009).
49. Baccin, C. *et al.* Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. *Nat. Cell Biol.* **22**, 38–48 (2020).
50. Pinho, S. & Frenette, P. S. Haematopoietic stem cell activity and interactions with the niche. *Nature Reviews Molecular Cell Biology* vol. 20 303–320 (2019).
51. Liggett, L. A. & Sankaran, V. G. Unraveling Hematopoiesis through the Lens of Genomics. *Cell* vol. 182 1384–1400 (2020).
52. Wilson, A. *et al.* Hematopoietic Stem Cells Reversibly Switch from Dormancy to Self-Renewal during Homeostasis and Repair. *Cell* **135**, 1118–1129 (2008).
53. King, K. Y. & Goodell, M. A. Inflammatory modulation of HSCs: Viewing the HSC as a foundation for the immune response. *Nature Reviews Immunology* vol. 11 685–692 (2011).
54. Takizawa, H., Boettcher, S. & Manz, M. G. Demand-adapted regulation of early hematopoiesis in infection and inflammation. *Blood* **119**, 2991–3002 (2012).
55. Christopher, M. J., Liu, F., Hilton, M. J., Long, F. & Link, D. C. Suppression of CXCL12 production by bone marrow osteoblasts is a common and critical pathway for cytokine-induced mobilization. *Blood* **114**, 1331–1339 (2009).
56. Semerad, C. L. *et al.* G-CSF potently inhibits osteoblast activity and CXCL12 mRNA expression in the bone marrow. *Blood* **106**, 3020–3027 (2005).
57. Kavgaci, H., Ozdemir, F., Aydin, F., Yavuz, A. & Yavuz, M. Endogenous granulocyte colony-stimulating factor (G-CSF) levels in chemotherapy-induced neutropenia and in neutropenia related with primary diseases. *J. Exp. Clin. Cancer Res.* **21**, 475–9 (2002).
58. Mehta, H. M., Malandra, M. & Corey, S. J. G-CSF and GM-CSF in Neutropenia. *J. Immunol.* **195**, 1341–1349 (2015).
59. Xia, J. *et al.* Somatic mutations and clonal hematopoiesis in congenital neutropenia. *Blood* **131**, 408–416 (2018).
60. Bernitz, J. M., Daniel, M. G., Fstkchyan, Y. S. & Moore, K. Granulocyte colony-stimulating factor mobilizes dormant hematopoietic stem cells without proliferation in mice. *Blood* **129**, 1901–1912 (2017).
61. Chen, X. *et al.* Bone Marrow Myeloid Cells Regulate Myeloid-Biased Hematopoietic Stem Cells via a Histamine-Dependent Feedback Loop. *Cell Stem Cell* **21**, 747-760.e7 (2017).

62. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).

63. Wang, N. D. *et al.* Impaired energy homeostasis in C/EBPα knockout mice. *Science (80-.).* **269**, 1108–1112 (1995).

64. Kühn, R., Schwenk, F., Aguet, M. & Rajewsky, K. Inducible gene targeting in mice. *Science (80-.).* **269**, 1427–1429 (1995).

65. Velasco-Hernandez, T., Säwén, P., Bryder, D. & Cammenga, J. Potential Pitfalls of the Mx1-Cre System: Implications for Experimental Modeling of Normal and Malignant Hematopoiesis. *Stem Cell Reports* **7**, 11–18 (2016).

66. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–30 (2015).

67. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

68. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat. Biotechnol.* **39**, 825–835 (2021).

69. Skokowa, J., Dale, D. C., Touw, I. P., Zeidler, C. & Welte, K. Severe congenital neutropenias. *Nat. Rev. Dis. Prim.* **3**, 17032 (2017).

70. Dror, Y. & Freedman, M. H. Shwachman-Diamond Syndrome: An Inherited Preleukemic Bone Marrow Failure Disorder With Aberrant Hematopoietic Progenitors and Faulty Marrow Microenvironment. *Blood* **94**, 3048–3054 (1999).

71. Bezzerri, V. & Cipolli, M. Shwachman-Diamond Syndrome: Molecular Mechanisms and Current Perspectives. *Molecular Diagnosis and Therapy* vol. 23 281–290 (2019).

72. Dror, Y. & Freedman, M. H. Shwachman-Diamond syndrome marrow cells show abnormally increased apoptosis mediated through the Fas pathway. *Blood* **97**, 3011–3016 (2001).

73. Kumar, S. *et al.* Repolarization of HSC attenuates HSCs failure in Shwachman–Diamond syndrome. *Leukemia* **35**, 1751–1762 (2021).

74. Pabst, T. *et al.* Dominant-negative mutations of CEBPA, encoding CCAAT/enhancer binding protein-α (C/EBPα), in acute myeloid leukemia. *Nat. Genet.* **27**, 263–270 (2001).

75. Wouters, B. J. *et al.* Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome. *Blood* **113**, 3088–3091 (2009).

76. Taskesen, E. *et al.* Prognostic impact, concurrent genetic mutations, and gene expression features of AML with CEBPA mutations in a cohort of 1182 cytogenetically normal AML patients: further evidence for CEBPA double mutant AML as a distinctive disease entity. *Blood* **117**, 2469–2475 (2011).

77. Avellino, R. & Delwel, R. Expression and regulation of C/EBPα in normal myelopoiesis and in malignant transformation. *Blood* vol. 129 2083–2091 (2017).

78. O'Reilly, E., Zeinabad, H. A. & Szegezdi, E. Hematopoietic versus leukemic stem cell quiescence: Challenges and therapeutic opportunities. *Blood Reviews* vol. 50 100850 (2021).

79. Wilson, M. *et al.* EVI1 interferes with myeloid maturation via transcriptional repression of Cebpa, via binding to two far downstream regulatory elements. *J. Biol. Chem.* **291**, 13591–13607 (2016).

80. Stengel, K. R., Ellis, J. D., Spielman, C. L., Bomber, M. L. & Hiebert, S. W. Definition of a small core transcriptional circuit regulated by AML1-ETO. *Mol. Cell* **81**, 530-545.e5 (2021).

81. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nature Reviews Molecular Cell Biology* vol. 16 144–154 (2015).

82. Herz, H. M. Enhancer deregulation in cancer and other diseases. *BioEssays* **38**, 1003–1015 (2016).
83. Northcott, P. A. *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428–434 (2014).
84. Francis, J. M. *et al.* EGFR variant heterogeneity in glioblastoma resolved through single-nucleus sequencing. *Cancer Discov.* **4**, 956–971 (2014).
85. Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science (80-.).* **351**, 1454–1458 (2016).
86. Hayday, A. C. *et al.* Activation of a translocated human c-myc gene by an enhancer in the immunoglobulin heavy-chain locus. *Nature* **307**, 334–340 (1984).
87. Taub, R. *et al.* Translocation of the c-myc gene into the immunoglobulin heavy chain locus in human Burkitt lymphoma and murine plasmacytoma cells. *Proc. Natl. Acad. Sci. U. S. A.* **79**, 7837–7841 (1982).
88. Cleary, M. L., Smith, S. D. & Sklar, J. Cloning and structural analysis of cDNAs for bcl-2 and a hybrid bcl-2/immunoglobulin transcript resulting from the t(14;18) translocation. *Cell* **47**, 19–28 (1986).
89. Khoury, J. D. *et al.* The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Myeloid and Histiocytic/Dendritic Neoplasms. *Leuk. 2022 367* **36**, 1703–1719 (2022).
90. Lugthart, S. *et al.* Clinical, molecular, and prognostic significance of WHO type inv(3)(q21q26.2)/t(3;3)(q21;q26.2) and various other 3q abnormalities in acute myeloid leukemia. *J. Clin. Oncol.* **28**, 3890–3898 (2010).
91. Gröschel, S. *et al.* Mutational spectrum of myeloid malignancies with inv(3)/t(3;3) reveals a predominant involvement of RAS/RTK signaling pathways. *Blood* **125**, 133–9 (2015).
92. Montefiori, L. E. *et al.* Enhancer hijacking drives oncogenic bcl11b expression in lineage-ambiguous stem cell leukemia. *Cancer Discov.* **11**, 2846–2867 (2021).
93. Mochizuki, N. *et al.* A novel gene, MEL1, mapped to 1p36.3 is highly homologous to the MDS1/EVI1 gene and is transcriptionally activated in t(1;3)(p36;q21)-positive leukemia cells. *Blood* **96**, 3209–3214 (2000).
94. Rubin, C. *et al.* t(3;21)(q26;q22): a recurring chromosomal abnormality in therapy- related myelodysplastic syndrome and acute myeloid leukemia. *Blood* **76**, 2594–2598 (1990).
95. Wang, T. *et al.* Ectopia associated MN1 fusions and aberrant activation in myeloid neoplasms with t(12;22)(p13;q12). *Cancer Gene Ther.* **27**, 810–818 (2020).
96. Scully, R., Panday, A., Elango, R. & Willis, N. A. DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nat. Rev. Mol. Cell Biol.* **20**, 698–714 (2019).
97. Tsai, A. G. & Lieber, M. R. Mechanisms of chromosomal rearrangement in the human genome. *BMC Genomics* **11**, S1 (2010).
98. Roy, R., Chun, J. & Powell, S. N. BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat. Rev. Cancer* **12**, 68–78 (2011).
99. Ghezraoui, H. *et al.* Chromosomal Translocations in Human Cells Are Generated by Canonical Nonhomologous End-Joining. *Mol. Cell* **55**, 829–842 (2014).
100. Robbiani, D. F. *et al.* AID Is Required for the Chromosomal Breaks in c-myc that Lead to c-myc/IgH Translocations. *Cell* **135**, 1028–1038 (2008).
101. Durkin, S. G. & Glover, T. W. Chromosome fragile sites. *Annual Review of Genetics* vol. 41 169–192 (2007).
102. Glover, T. W., Wilson, T. E. & Arlt, M. F. Fragile sites in cancer: More than meets the eye. *Nature Reviews Cancer* vol. 17 489–501 (2017).
103. Zlotorynski, E. *et al.* Molecular Basis for Expression of Common and Rare Fragile Sites. *Mol. Cell. Biol.* **23**, 7143–7151 (2003).
104. Bacolla, A., Tainer, J. A., Vasquez, K. M. & Cooper, D. N. Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences. *Nucleic Acids Res.* **44**, 5673–5688 (2016).

105. Letessier, A. *et al.* Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. *Nature* **470**, 120–124 (2011).

106. Le Beau, M. M. *et al.* Replication of a common fragile site, FRA3B, occurs late in S phase and is delayed further upon induction: Implications for the mechanism of fragile site induction. *Hum. Mol. Genet.* **7**, 755–761 (1998).

107. Mizuno, K., Miyabe, I., Schalbetter, S. A., Carr, A. M. & Murray, J. M. Recombination-restarted replication makes inverted chromosome fusions at inverted repeats. *Nature* **493**, 246–249 (2013).

108. Fungtammasan, A., Walsh, E., Chiaromonte, F., Eckert, K. A. & Makova, K. D. A genome-wide analysis of common fragile sites: What features determine chromosomal instability in the human genome? *Genome Res.* **22**, 993–1005 (2012).

109. Helmrich, A., Ballarino, M. & Tora, L. Collisions between Replication and Transcription Complexes Cause Common Fragile Site Instability at the Longest Human Genes. *Mol. Cell* **44**, 966–977 (2011).

110. Wilson, T. E. *et al.* Large transcription units unify copy number variants and common fragile sites arising under replication stress. *Genome Res.* **25**, 189–200 (2015).

111. Canela, A. *et al.* Genome Organization Drives Chromosome Fragility. *Cell* **170**, 507-521.e18 (2017).

112. Gómez-Herreros, F. DNA Double Strand Breaks and Chromosomal Translocations Induced by DNA Topoisomerase II. *Frontiers in Molecular Biosciences* vol. 6 141 (2019).

113. Zhang, Y. *et al.* Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* **148**, 908–21 (2012).

114. Hsieh, T. H. S. *et al.* Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162**, 108–119 (2015).

115. Morishita, K., Parganas, E., Matsugi, T. & Ihle, J. N. Expression of the Evi-1 zinc finger gene in 32Dc13 myeloid cells blocks granulocytic differentiation in response to granulocyte colony-stimulating factor. *Mol. Cell. Biol.* **12**, 183–189 (1992).

116. Goyama, S. *et al.* Evi-1 Is a Critical Regulator for Hematopoietic Stem Cells and Transformed Leukemic Cells. *Cell Stem Cell* **3**, 207–220 (2008).

117. Laricchia-Robbio, L. & Nucifora, G. Significant increase of self-renewal in hematopoietic cells after forced expression of EVI1. *Blood Cells, Mol. Dis.* **40**, 141–147 (2008).

118. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).

119. Wilson, N. K. *et al.* Combinatorial transcriptional control in blood stem/progenitor cells: Genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* **7**, 532–544 (2010).

120. Ryu, J., Kim, H., Yang, D., Lee, A. J. & Jung, I. A new class of constitutively active super-enhancers is associated with fast recovery of 3D chromatin loops. *BMC Bioinformatics* **20**, 25–36 (2019).

121. van Arensbergen, J., van Steensel, B. & Bussemaker, H. J. In search of the determinants of enhancer-promoter interaction specificity. *Trends in Cell Biology* vol. 24 695–702 (2014).

122. Schmidt, D. *et al.* A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res.* **20**, 578 (2010).

123. Abboud, N. *et al.* A cohesin-OCT4 complex mediates Sox enhancers to prime an early embryonic lineage. *Nat. Commun.* **6**, 1–14 (2015).

124. Weintraub, A. S. *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* **171**, 1573-1588.e28 (2017).

125. Butler, J. E. F. & Kadonaga, J. T. Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes Dev.* **15**, 2515–2519 (2001).

126. Juven-Gershon, T., Hsu, J. Y. & Kadonaga, J. T. Caudal, a key developmental regulator, is a DPE-specific transcriptional factor. *Genes Dev.* **22**, 2823–2830 (2008).

127. Martinez-Ara, M., Comoglio, F., van Arensbergen, J. & van Steensel, B. Systematic analysis of intrinsic enhancer-promoter compatibility in the mouse genome. *Mol. Cell* 2021.10.21.465269 (2022) doi:10.1016/j.molcel.2022.04.009.

128. Tsui, C. *et al.* DCas9-targeted locus-specific protein isolation method identifies histone gene regulators. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E2734–E2741 (2018).

129. Liu, X. *et al.* In Situ Capture of Chromatin Interactions by Biotinylated dCas9. *Cell* **170**, 1028-1043.e19 (2017).

130. Tsai, F. Y. *et al.* An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature* **371**, 221–226 (1994).

131. de Pater, E. *et al.* Gata2 is required for HSC generation and survival. *J. Exp. Med.* **210**, 2843–2850 (2013).

132. Menendez-Gonzalez, J. B. *et al.* Gata2 as a Crucial Regulator of Stem Cells in Adult Hematopoiesis and Acute Myeloid Leukemia. *Stem Cell Reports* **13**, 291–306 (2019).

133. Ostergaard, P. *et al.* Mutations in GATA2 cause primary lymphedema associated with a predisposition to acute myeloid leukemia (Emberger syndrome). *Nat. Genet.* **43**, 929–931 (2011).

134. Hsu, A. P. *et al.* GATA2 haploinsufficiency caused by mutations in a conserved intronic element leads to MonoMAC syndrome. *Blood* **121**, 3830–3837 (2013).

135. Dickinson, R. E. *et al.* Exome sequencing identifies GATA-2 mutation as the cause of dendritic cell, monocyte, B and NK lymphoid deficiency. *Blood* **118**, 2656–2658 (2011).

136. Hsu, A. P. *et al.* Mutations in GATA2 are associated with the autosomal dominant and sporadic monocytopenia and mycobacterial infection (MonoMAC) syndrome. *Blood* **118**, 2653–2655 (2011).

137. Hahn, C. N. *et al.* Heritable GATA2 mutations associated with familial myelodysplastic syndrome and acute myeloid leukemia. *Nat. Genet.* **43**, 1012–1019 (2011).

138. Rodrigues, N. P. *et al.* Haploinsufficiency of GATA-2 perturbs adult hematopoietic stem-cell homeostasis. *Blood* **106**, 477–484 (2005).

139. Rodrigues, N. P. *et al.* GATA-2 regulates granulocyte-macrophage progenitor cell function. *Blood* **112**, 4862–4873 (2008).

140. Katayama, S. *et al.* GATA2 haploinsufficiency accelerates EVI1-driven leukemogenesis. *Blood* **130**, 908–919 (2017).

141. Wilson, A. *et al.* c-Myc controls the balance between hematopoietic stem cell self-renewal and differentiation. *Genes Dev.* **18**, 2747–2763 (2004).

142. Scheicher, R. *et al.* CDK6 as a key regulator of hematopoietic and leukemic stem cell activation. *Blood* **125**, 90–101 (2015).

143. Sun, Y. *et al.* Safety and Efficacy of Bromodomain and Extra-Terminal Inhibitors for the Treatment of Hematological Malignancies and Solid Tumors: A Systematic Study of Clinical Trials. *Front. Pharmacol.* **11**, 621093 (2020).

144. Reiter, F., Wienerroither, S. & Stark, A. Combinatorial function of transcription factors and cofactors. *Current Opinion in Genetics and Development* vol. 43 73–81 (2017).

145. Mansour, M. R. *et al.* An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science (80-.).* **346**, 1373–1377 (2014).

146. Abraham, B. J. *et al.* Small genomic insertions form enhancers that misregulate oncogenes. *Nat. Commun.* **8**, 1–13 (2017).

147. Segert, J. A., Gisselbrecht, S. S. & Bulyk, M. L. Transcriptional Silencers: Driving Gene Expression with the Brakes On. *Trends in Genetics* vol. 37 514–527 (2021).

148. Cai, Y. *et al.* H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. *Nat. Commun.* **12**, 1–22 (2021).

149. Lugthart, S. *et al.* High EVI1 levels predict adverse outcome in acute myeloid leukemia: Prevalence of EVI1 overexpression and chromosome 3q26 abnormalities underestimated. *Blood* **111**, 4329–4337 (2008).

150. Ogawa, S. *et al.* Abnormal expression of Evi-1 gene in human leukemias. *Hum. Cell* **9**, 323–32 (1996).

151. Mitani, K. *et al.* Generation of the AML1-EVI-1 fusion gene in the t(3;21)(q26;q22) causes blastic crisis in chronic myelocytic leukemia. *EMBO J.* **13**, 504 (1994).

152. Bindels, E. M. J. *et al.* EVI1 is critical for the pathogenesis of a subset of MLL-AF9-rearranged AMLs. *Blood* **119**, 5838–5849 (2012).

153. Okada, Y. *et al.* hDOT1L links histone methylation to leukemogenesis. *Cell* **121**, 167–178 (2005).

154. Feng, Q. *et al.* Methylation of H3-Lysine 79 Is Mediated by a New Family of HMTases without a SET Domain. *Curr. Biol.* **12**, 1052–1058 (2002).

155. Van Leeuwen, F., Gafken, P. R. & Gottschling, D. E. Dot1p modulates silencing in yeast by methylation of the nucleosome core. *Cell* **109**, 745–756 (2002).

156. Krivtsov, A. V. *et al.* Cell of origin determines clinically relevant subtypes of MLL-rearranged AML. *Leukemia* **27**, 852–860 (2013).

157. Zhang, Y. *et al.* Mds1CreERT2, an inducible Cre allele specific to adult-repopulating hematopoietic stem cells. *Cell Rep.* **36**, (2021).

158. Wimmer, K., Vinatzer, U., Zwirn, P., Fonatsch, C. & Wieser, R. Comparative expression analysis of the antagonistic transcription factors EVI1 and MDS1-EVI1 in murine tissues and during in vitro hematopoietic differentiation. *Biochem. Biophys. Res. Commun.* **252**, 691–696 (1998).

159. Zhang, Y. *et al.* PR-domain - containing Mds1-Evi1 is critical for long-term hematopoietic stem cell function. *Blood* **118**, 3853–3861 (2011).

160. Soderholm, J., Kobayashi, H., Mathieu, C., Rowley, J. D. & Nucifora, G. The leukemia-associated gene MDS1/EVI1 is a new type of GATA-binding transactivator. *Leukemia* **11**, 352–358 (1997).

161. Sitailo, S., Sood, R., Barton, K. & Nucifora, G. Forced expression of the leukemia-associated gene EVI1 in ES cells: A model for myeloid leukemia with 3q26 rearrangements. *Leukemia* vol. 13 1639–1645 (1999).

162. Sood, R., Talwar-Trikha, A., Chakrabarti, S. R. & Nucifora, G. MDS1/EVI1 enhances TGF-β1 signaling and strengthens its growth-inhibitory effect, but the leukemia-associated fusion protein AML1/MDS1/EVI1, product of the t(3;21), abrogates growth-inhibition in response to TGF-β1. *Leukemia* **13**, 348–357 (1999).

163. Nitta, E. *et al.* Oligomerization of Evi-1 regulated by the PR domain contributes to recruitment of corepressor CtBP. *Oncogene* **24**, 6165–6173 (2005).

164. Kataoka, K. *et al.* Evi1 is essential for hematopoietic stem cell self-renewal, and its expression marks hematopoietic cells with long-term multilineage repopulating activity. *J. Exp. Med.* **208**, 2403–2416 (2011).

165. Milne, T. A. Mouse models of MLL leukemia: recapitulating the human disease. *Blood* **129**, 2217–2223 (2017).

166. Nanjundan, M. *et al.* Amplification of MDS1/EVI1 and EVI1, located in the 3q26.2 amplicon, is associated with favorable patient prognosis in ovarian cancer. *Cancer Res.* **67**, 3074–3084 (2007).

167. Volkert, S. *et al.* Amplification of EVI1 on cytogenetically cryptic double minutes as new mechanism for increased expression of EVI1. *Cancer Genet.* **207**, 103–108 (2014).

168. Jones, P. A. & Baylin, S. B. The Epigenomics of Cancer. *Cell* vol. 128 683–692 (2007).

169. Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science (80-.).* **339**, 957–959 (2013).

170. Thorn, J., Molloy, P. & Iland, H. SSCP detection of N-ras promoter mutations in AML patients. *Exp. Hematol.* **23**, 1098–103 (1995).

171. Bookstein, R. *et al.* Promoter deletion and loss of retinoblastoma gene expression in human prostate carcinoma. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 7762–7766 (1990).

172. Shi, J. *et al.* Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation. *Genes Dev.* **27**, 2648–2662 (2013).

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

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

175. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

176. Doolittle, W. F. & Brunet, T. D. P. On causal roles and selected effects: Our genome is mostly junk. *BMC Biol.* **15**, 1–9 (2017).

177. Ohno, S. So much 'junk' DNA in our genome. *Brookhaven Symp. Biol.* **23**, 366–70 (1972).

178. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).

179. Ponting, C. P. & Hardison, R. C. What fraction of the human genome is functional? *Genome Res.* **21**, 1769–76 (2011).

180. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).

181. Rahman, S. & Mansour, M. R. The role of noncoding mutations in blood cancers. *DMM Disease Models and Mechanisms* vol. 12 (2019).

182. Adli, M. The CRISPR tool kit for genome editing and beyond. *Nat. Commun.* **9**, 1–13 (2018).

183. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–83 (2013).

184. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442–51 (2013).

185. Nuñez, J. K. *et al.* Genome-wide programmable transcriptional memory by CRISPR-based epigenome editing. *Cell* **184**, 2503-2519.e17 (2021).

186. Evans, T., Reitman, M. & Felsenfeld, G. An erythrocyte-specific DNA-binding factor recognizes a regulatory sequence common to all chicken globin genes. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 5976–5980 (1988).

187. Martin, D. I. K., Tsai, S. F. & Orkin, S. H. Increased γ-globin expression in a nondeletion HPFH mediated by an erythroid-specific DNA-binding factor. *Nature* **338**, 435–438 (1989).

188. Mantovani, R. *et al.* An erythroid specific nuclear factor binding to the proximal CACCC box of the β-globin gene promoter. *Nucleic Acids Res.* **16**, 4299–4313 (1988).

189. Wall, L., DeBoer, E. & Grosveld, F. The human beta-globin gene 3' enhancer contains multiple binding sites for an erythroid-specific protein. *Genes Dev.* **2**, 1089–1100 (1988).

190. Gumucio, D. L. *et al.* Nuclear proteins that bind the human gamma-globin gene promoter: alterations in binding produced by point mutations associated with hereditary persistence of fetal hemoglobin. *Mol. Cell. Biol.* **8**, 5310–5322 (1988).

191. Tsai, S. F. *et al.* Cloning of cDNA for the major DNA-binding protein of the erythroid lineage through expression in mammalian cells. *Nature* **339**, 446–51 (1989).

192. Evans, T. & Felsenfeld, G. The erythroid-specific transcription factor Eryf1: a new finger protein. *Cell* **58**, 877–85 (1989).

193. Tsang, A. P. *et al.* FOG, a multitype zinc finger protein, acts as a cofactor for transcription factor GATA-1 in erythroid and megakaryocytic differentiation. *Cell* **90**, 109–119 (1997).

194. Yamamoto, M. *et al.* Activity and tissue-specific expression of the transcription factor NF-E1 multigene family. *Genes Dev.* **4**, 1650–1662 (1990).

195. Ho, I. C. *et al.* Human GATA-3: a lineage-restricted transcription factor that regulates the expression of the T cell receptor alpha gene. *EMBO J.* **10**, 1187–1192 (1991).

196. Dorfman, D. M., Wilson, D. B., Bruns, G. A. P. & Orkin, S. H. Human transcription factor GATA-2: Evidence for regulation of preproendothelin-1 gene expression in endothelial cells. *J. Biol. Chem.* **267**, 1279–1285 (1992).

197. Bresnick, E. H., Katsumura, K. R., Lee, H. Y., Johnson, K. D. & Perkins, A. S. Master regulatory GATA transcription factors: Mechanistic principles and emerging links to hematologic malignancies. *Nucleic Acids Research* vol. 40 5819–5831 (2012).

198. Grass, J. A. *et al.* 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 (2003).

199. Tsai, F.-Y. & Orkin, S. H. Transcription Factor GATA-2 Is Required for Proliferation/Survival of Early Hematopoietic Cells and Mast Cell Formation, But Not for Erythroid and Myeloid Terminal Differentiation. *Blood* **89**, 3636–3643 (1997).

200. Ling, K. W. *et al.* GATA-2 plays two functionally distinct roles during the ontogeny of hematopoietic stem cells. *J. Exp. Med.* **200**, 871–882 (2004).

201. Nagai, T. *et al.* Transcription factor GATA-2 is expressed in erythroid, early myeloid, and CD34+ human leukemia-derived cell lines. *Blood* **84**, 1074–1084 (1994).

202. Bresnick, E. H. & Johnson, K. D. Blood disease–causing and –suppressing transcriptional enhancers: General principles and GATA2 mechanisms. *Blood Advances* vol. 3 2045–2056 (2019).

203. Sanalkumar, R. *et al.* Mechanism governing a stem cell-generating cis-regulatory element. *Proc. Natl. Acad. Sci. U. S. A.* **111**, (2014).

204. Johnson, K. D. *et al.* Cis-element mutated in GATA2-dependent immunodeficiency governs hematopoiesis and vascular integrity. *J. Clin. Invest.* **122**, 3692–3704 (2012).

205. Johnson, K. D. *et al.* Cis-regulatory mechanisms governing stem and progenitor cell transitions. *Sci. Adv.* **1**, (2015).

206. Hirabayashi, S., Wlodarski, M. W., Kozyra, E. & Niemeyer, C. M. Heterogeneity of GATA2-related myeloid neoplasms. *International Journal of Hematology* vol. 106 175–182 (2017).

207. Persons, D. A. *et al.* Enforced expression of the GATA-2 transcription factor blocks normal hematopoiesis. *Blood* **93**, 488–499 (1999).

208. Tipping, A. J. *et al.* High GATA-2 expression inhibits human hematopoietic stem and progenitor cell function by effects on cell cycle. *Blood* **113**, 2661–2672 (2009).

209. Nandakumar, S. K. *et al.* Low-level GATA2 overexpression promotes myeloid progenitor self-renewal and blocks lymphoid differentiation in mice. *Exp. Hematol.* **43**, 565-577.e10 (2015).

210. Vicente, C. *et al.* Overexpression of GATA2 predicts an adverse prognosis for patients with acute myeloid leukemia and it is associated with distinct molecular abnormalities. *Leukemia* vol. 26 550–554 (2012).

211. Luesink, M. *et al.* High GATA2 expression is a poor prognostic marker in pediatric acute myeloid leukemia. *Blood* **120**, 2064–2075 (2012).

212. The Cancer Genome Atlas Research Network *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–74 (2013).

213. Tyner, J. W. *et al.* Functional genomic landscape of acute myeloid leukaemia. *Nature* **562**, 526–531 (2018).

214. Leubolt, G., Redondo Monte, E. & Greif, P. A. GATA2 mutations in myeloid malignancies: Two zinc fingers in many pies. *IUBMB Life* vol. 72 151–158 (2020).

215. Tien, F. M. *et al.* GATA2 zinc finger 1 mutations are associated with distinct clinico-biological features and outcomes different from GATA2 zinc finger 2 mutations in adult acute myeloid leukemia. *Blood Cancer J.* **8**, 87 (2018).

216. Zhang, S. J., Shi, J. Y. & Li, J. Y. GATA-2 L359 V mutation is exclusively associated with CML progression but not other hematological malignancies and GATA-2 P250A is a novel single nucleotide polymorphism. *Leuk. Res.* **33**, 1141–1143 (2009).

217. Katsumura, K. R. *et al.* Human leukemia mutations corrupt but do not abrogate GATA-2 function. *Proc. Natl. Acad. Sci.* **115**, E10109–E10118 (2018).

218. Chong, C. E. *et al.* Differential effects on gene transcription and hematopoietic differentiation correlate with GATA2 mutant disease phenotypes. *Leukemia* **32**, 194–202 (2018).

219. Iwasaki, H. *et al.* The order of expression of transcription factors directs hierarchical specification of hematopoietic lineages. *Genes Dev.* **20**, 3010–3021 (2006).

220. Saida, S. *et al.* Gata2 deficiency delays leukemogenesis while contributing to aggressive leukemia phenotype in Cbfb-MYH11 knockin mice. *Leukemia* **34**, 759–770 (2020).

221. Di Genua, C. *et al.* C/EBPα and GATA-2 Mutations Induce Bilineage Acute Erythroid Leukemia through Transformation of a Neomorphic Neutrophil-Erythroid Progenitor. *Cancer Cell* **37**, 690-704.e8 (2020).

222. Smith, M. L., Cavenagh, J. D., Lister, T. A. & Fitzgibbon, J. Mutation of CEBPA in Familial Acute Myeloid Leukemia. *N. Engl. J. Med.* **351**, 2403–2407 (2004).

223. Celton, M. *et al.* Epigenetic regulation of GATA2 and its impact on normal karyotype acute myeloid leukemia. *Leukemia* **28**, 1617–1626 (2014).

224. Al Seraihi, A. F. *et al.* GATA2 monoallelic expression underlies reduced penetrance in inherited GATA2-mutated MDS/AML. *Leukemia* **32**, 2502–2507 (2018).

225. Hou, H. A. *et al.* WT1 mutation in 470 adult patients with acute myeloid leukemia: Stability during disease evolution and implication of its incorporation into a survival scoring system. *Blood* **115**, 5222–5231 (2010).

226. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).

227. Huang, T., Sun, L., Yuan, X. & Qiu, H. Thrombospondin-1 is a multifaceted player in tumor progression. *Oncotarget* **8**, 84546–84558 (2017).

228. Jaiswal, S. *et al.* CD47 Is Upregulated on Circulating Hematopoietic Stem Cells and Leukemia Cells to Avoid Phagocytosis. *Cell* **138**, 271–285 (2009).

229. Majeti, R. *et al.* CD47 Is an Adverse Prognostic Factor and Therapeutic Antibody Target on Human Acute Myeloid Leukemia Stem Cells. *Cell* **138**, 286–299 (2009).

230. Maute, R., Xu, J. & Weissman, I. L. CD47-SIRPα-targeted therapeutics: status and prospects. *Immuno-Oncology Technol.* **13**, 100070 (2022).

231. Zhu, L. *et al.* THBS1 Is a Novel Serum Prognostic Factors of Acute Myeloid Leukemia. *Front. Oncol.* **9**, 1567 (2019).

232. Kahoul, Y. *et al.* Emerging Roles for the INK4a/ARF (CDKN2A) Locus in Adipose Tissue: Implications for Obesity and Type 2 Diabetes. *Biomolecules* **10**, 1–16 (2020).

233. Jacobs, J. L., Kieboom, K., Marino, S., DePinho, R. A. & Van Lohuizen, M. The oncogene and Polycombgroup gene bmi-1 regulates cell proliferation and senescence through the ink4a locus. *Nature* **397**, 164–168 (1999).

234. Tanaka, S. *et al.* Ezh2 augments leukemogenicity by reinforcing differentiation blockage in acute myeloid leukemia. *Blood* **120**, 1107–1117 (2012).

235. Chaturvedi, A. *et al.* Mutant IDH1 promotes leukemogenesis in vivo and can be specifically targeted in human AML. *Blood* **122**, 2877–2887 (2013).

236. De Jonge, H. J. M. *et al.* AML at older age: Age-related gene expression profiles reveal a paradoxical down-regulation of p16INK4A mRNA with prognostic significance. *Blood* **114**, 2869–2877 (2009).

237. Yáñez, A., Ng, M. Y., Hassanzadeh-Kiabi, N. & Goodridge, H. S. IRF8 acts in lineage-committed rather than oligopotent progenitors to control neutrophil vs monocyte production. *Blood* **125**, 1452–1459 (2015).

238. Holtschke, T. *et al.* Immunodeficiency and chronic myelogenous leukemia-like syndrome in mice with a targeted mutation of the ICSBP gene. *Cell* **87**, 307–317 (1996).

239. Pogosova-Agadjanyan, E. L. *et al.* The Prognostic Significance of IRF8 Transcripts in Adult Patients with Acute Myeloid Leukemia. *PLoS One* **8**, (2013).

240. Silva, F. P. G. *et al.* Gene expression profiling of minimally differentiated acute myeloid leukemia: M0 is a distinct entity subdivided by RUNX1 mutation status. *Blood* **114**, 3001–3007 (2009).

241. Cao, Z. *et al.* ZMYND8-regulated IRF8 transcription axis is an acute myeloid leukemia dependency. *Mol. Cell* **81**, 3604-3622.e10 (2021).

242. Unnisa, Z. *et al.* Meis1 preserves hematopoietic stem cells in mice by limiting oxidative stress. *Blood* **120**, 4973–4981 (2012).

243. Pineault, N., Helgason, C. D., Lawrence, H. J. & Humphries, R. K. Differential expression of Hox, Meis1, and Pbx1 genes in primitive cells throughout murine hematopoietic ontogeny. *Exp. Hematol.* **30**, 49–57 (2002).

244. Verhaak, R. G. W. *et al.* Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): Association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. *Blood* **106**, 3747–3754 (2005).

245. Armstrong, S. A. *et al.* MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.* **30**, 41–47 (2002).

246. Lawrence, H. J. *et al.* Frequent co-expression of the HOXA9 and MEIS1 homeobox genes in human myeloid leukemias. *Leukemia* **13**, 1993–1999 (1999).

247. Kroon, E. *et al.* Hoxa9 transforms primary bone marrow cells through specific collaboration with Meis1a but not Pbx1b. *EMBO J.* **17**, 3714–3725 (1998).

248. Zeisig, B. B. *et al.* Hoxa9 and Meis1 Are Key Targets for MLL-ENL-Mediated Cellular Immortalization. *Mol. Cell. Biol.* **24**, 617–628 (2004).

249. Muranyi, A. *et al.* Npm1 Haploinsufficiency in collaboration with MEIS1 is sufficient to induce AML in mice. *Blood Adv.* (2022) doi:10.1182/bloodadvances.2022007015.

250. Collins, C. *et al.* C/EBPα is an essential collaborator in Hoxa9/Meis1-mediated leukemogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 9899–9904 (2014).

251. Bi, X. *et al.* Deletion of Irf5 protects hematopoietic stem cells from DNA damage-induced apoptosis and suppresses γ-irradiation-induced thymic lymphomagenesis. *Oncogene* **33**, 3288–3297 (2014).

252. Thalhammer-Scherrer, R. *et al.* The immunophenotype of 325 adult acute leukemias: Relationship to morphologic and molecular classification and proposal for a minimal screening program highly predictive for lineage discrimination. *Am. J. Clin. Pathol.* **117**, 380–389 (2002).

253. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**, D845–D855 (2020).

254. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).

255. Bock, C. *et al.* DNA Methylation Dynamics during In Vivo Differentiation of Blood and Skin Stem Cells. *Mol. Cell* **47**, 633–647 (2012).

256. Farlik, M. *et al.* DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. *Cell Stem Cell* **19**, 808–822 (2016).

257. Bröske, A. M. *et al.* DNA methylation protects hematopoietic stem cell multipotency from myeloerythroid restriction. *Nat. Genet.* **41**, 1207–1215 (2009).

258. Cole, C. B. *et al.* Haploinsufficiency for DNA methyltransferase 3A predisposes hematopoietic cells to myeloid malignancies. *J. Clin. Invest.* **127**, 3657–3674 (2017).

259. Hidalgo, I. *et al.* Ezh1 is required for hematopoietic stem cell maintenance and prevents senescence-like cell cycle arrest. *Cell Stem Cell* **11**, 649–662 (2012).

260. Mochizuki-Kashio, M. *et al.* Dependency on the polycomb gene Ezh2 distinguishes fetal from adult hematopoietic stem cells. *Blood* **118**, 6553–6561 (2011).

261. Xie, H., Ye, M., Feng, R. & Graf, T. Stepwise reprogramming of B cells into macrophages. *Cell* **117**, 663–676 (2004).

262. Martens, J. H. A. & Stunnenberg, H. G. BLUEPRINT: mapping human blood cell epigenomes. *Haematologica* **98**, 1487–9 (2013).

263. Cooper, S., Guo, H. & Friedman, A. D. The +37 kb Cebpa enhancer is critical for Cebpa myeloid gene expression and contains functional sites that bind SCL, GATA2, C/EBPα, PU.1, and additional Ets factors. *PLoS One* **10**, e0126385 (2015).

264. Valk, P. J. M. *et al.* Prognostically Useful Gene-Expression Profiles in Acute Myeloid Leukemia. *N. Engl. J. Med.* **350**, 1617–1628 (2004).

265. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology* vol. 20 590–607 (2019).

266. Lin, F. T., MacDougald, O. A., Diehl, A. M. & Lane, M. D. A 30-kDa alternative translation product of the CCAAT/enhancer binding protein α message: Transcriptional activator lacking antimitotic activity. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 9606–9610 (1993).

267. Figueroa, M. E. *et al.* DNA Methylation Signatures Identify Biologically Distinct Subtypes in Acute Myeloid Leukemia. *Cancer Cell* **17**, 13–27 (2010).

268. El-Sharkawi, D. *et al.* Variable outcome and methylation status according to CEBPA mutant type in double-mutated acute myeloid leukemia patients and the possible implications for treatment. *Haematologica* **103**, 91–100 (2018).

A

Addendum

NEDERLANDSE SAMENVATTING

Hematopoëse is de vorming van bloedcellen en bloedplaatjes door de differentiatie van hematopoëitische stamcellen. Hierbij is het aansturen van het juiste transcriptionele programma tijdens dit differentiatieproces door epigenetische mechanismen van essentieel belang. Verstoringen in deze mechanismen leiden tot deregulatie van genen die van belang zijn bij de proliferatie of differentiatie, wat uiteindelijk kan leidden tot leukemie. Het belang van epigenetische factoren, blijkt uit het feit dat bijna 75% van de acute myeloïde leukemie (AML)-patiënten mutaties in epigenetische regulatoren heeft. De term epigenetica is geïntroduceerd door Waddington, die epigenetica beschreef als een landschapsmodel waarin cellen werden voorgesteld als knikkers die zich tijdens de ontwikkeling specialiseren door in verschillende groeven van een heuvel af te rollen. Inmiddels begrijpen we steeds beter hoe epigenetica werkt. Tegenwoordig weten we dat epigenetica bestaat uit meetbare gegevens zoals DNA-methylatie, histon-acetylatie en binding van transcriptiefactoren. Door de komst van next generation sequencing (NGS) technieken kan de moleculair bioloog tegenwoordig het epigenetische landschap van een heel genoom in kaart brengen, zelfs tot op het niveau van een enkele cel. Echter, deze ongekende hoeveelheid aan nieuwe data brengt nieuwe uitdagingen met zich mee en dan vooral op het gebied van data-analyse en de interpretatie ervan. Bioinformatica is tegenwoordig een onmisbare discipline om deze uitdagingen aan te gaan. Het onderzoek in dit proefschrift beschrijft zowel de moleculaire biologie alsmede de bioinformatica om de rol van transcriptiefactoren in hematopoëse te begrijpen, mechanismen van enhancer-herpositionering in leukemie te bestuderen en epigenetische processen in leukemogenese te onderzoeken.

In **hoofdstuk 2** bestuderen we het belang van C/EBPA voor de myeloïde differentiatie en het onderhouden van de HSC-populatie met behulp van een muizenmodel waarbij de +37 kb *Cebpa* enhancer (+42 kb bij mensen) is gedeleteerd. Eerder werk liet zien dat de deletie van het *Cebpa* gen[1] of zijn hematopoëitische enhancer leidt tot neutropenie met depletie van het LT-HSC compartiment tot gevolg. Alhoewel eerder werd geconcludeerd dat dit laatste een gevolg is van een directe rol van C/EBPA in LT-HSC functie, laten we in dit proefschrift zien dat dit in feite het gevolg is van een cel-extrinsieke gebeurtenis is, die geactiveerd wordt door neutropenie. Deze conclusie wordt ondersteund door vier verschillende feiten: *Cebpa* was nauwelijks aantoonbaar in LT-HSCs in single-cel transcriptoom studies, het verlies van de LT-HSC cellen was proportioneel aan de mate van neutropenie, lymfocyten werden nog wel gemaakt in muizen zonder *Cebpa* en als laatste secundaire transplantaties van de 37-kb gedeleteerde beenmergcellen leidde ook tot neutropenie.

Het werk beschreven in **hoofdstuk 3 tot en met 5** gaat over AML met 3q26 herschikkingen. Herpositionering van de *GATA2* hematopoëitische super-enhancer (SE) naar het *MECOM* locus op 3q26 stuurt de over-expressie van het *EVI1* isoform aan in AML met inv(3)/t(3;3), met vermindering van *GATA2* expressie als gevolg[4,5]. Het transcriptionele isoform *MDS1-EVI1*, die

ook afgeschreven wordt van het *MECOM*-locus, komt niet tot expressie bij deze leukemieën. In **hoofdstuk 3** tonen we aan dat een vergelijkbaar mechanisme aanwezig is in AML met atypische 3q26 herschikkingen, waarbij elke keer een SE in de nabijheid van *EVI1* wordt gebracht wat resulteert in *EVI1* over-expressie. Tevens bevatten deze leukemieën ook geen *MDS1-EVI1* expressie en hebben ze vaak mono-allelische *GATA2* expressie. Concluderend, de AMLs met 3q26 herschikkingen kunnen beschouwd worden als één type leukemie met gedeelde pathobiologische eigenschappen. **Hoofdstuk 4** onderzoekt het mechanisme van AML met t(3;8), waarbij een SE van *MYC* nabij *EVI1* wordt gepositioneerd. Om dit te bestuderen, werd door middel van CRISPR-Cas9 een op een patiënt gebaseerd t(3;8) model in K562 cellen gegenereerd. Dit model bevat *eGFP* achter *EVI1* (*EVI1-eGFP*), waardoor we de mogelijk hebben *EVI1* expressie te volgen indien we verschillende functionele domeinen van de SE regio verwijderen. Deze strategie leidde tot de ontdekking van een hematopoëtische module van de MYC SE die essentieel is voor *EVI1* expressie alsmede de promotor-enhancer interactie. Deze interactie was afhankelijk van CTCF binding aan de *MYC* SE en ook de binding aan de promotor van *EVI1*. Deze CTCF bindingsplaats nabij de promotor van *EVI1* bleek altijd behouden te blijven bij andere 3q26 herschikte leukemieën, wat suggereert dat deze van belang is bij *EVI1* herschikkingen. In **hoofdstuk 5** hebben we motieven in de herschikte *GATA2*-enhancer geïdentificeerd die essentieel zijn voor de expressie van *EVI1*. Dit hebben we gedaan met behulp van een op CRISPR-Cas9 gebaseerde enhancer screen. Deze screen bestond uit een lentivirale collectie van 3,239 sgRNAs. Cellen met verminderde GFP/EVI1 expressie waren verrijkt voor sgRNAs voor een p300-interactie regio, waar onder andere heptad hematopoëtische TFs binden. Mutaties in de MYB bindingsplaats van deze regio resulteerde in verminderde *EVI1* expressie terwijl *GATA2* expressie onveranderd bleef. Farmacologische inhibitie van MYB had hetzelfde effect, wat suggereert dat dit een manier is om selectief te interfereren met oncogene activatie van *EVI1*.

Studies in **hoofdstuk 6 en 7** onderzoeken mechanismen van epigenetische deregulatie in AML en de rol hiervan in leukemogenese. **In hoofdstuk 6** hebben we whole genome exome sequencing (WES) en RNA-seq data van 200 AMLs geanalyseerd om allel-specifieke expressie (ASE) te kunnen detecteren, wat een aanwijzing zou kunnen zijn voor veranderingen in cis-regulatie regio's. Deze analyse detecteerde frequent ASE van *GATA2* in AML en dan met name in patiënten met *CEBPA* dubbel mutaties (*CEBPAdm*). Veel van deze *CEBPAdm* bevatten ook *GATA2* mutaties waarbij het gemuteerde allel altijd het allel was dat het meeste tot expressie kwam, overeenkomstig met bevindingen van eerder onderzoek [6]. Verder toonden we aan dat *GATA2* ASE somatisch is en afwezig bij remissie. Het is afkomstig van allel-specifieke methylatie van de promotor met hyperactivatie van de enhancer van het andere allel tot gevolg. Een mogelijke verklaring is dat dit een voorbeeld is van een epimutatie die samen met een genetische mutatie leukemogenese aanstuurt. Het gevolg van methylatie voor de regulatie van myeloïde TFs is verder onderzocht in **hoofdstuk 7**. Voorafgaand werk van Wouters et al. identificeerde een subgroep van leukemieën met wijdverspreide methylatie

resulterend in *CEBPA* silencing met een myeloïde/lymfatisch fenotype[7,8]. Onafhankelijk hiervan rapporteerde Gebhard et al een gelijke subgroep met een CpG island methylatie fenotype (CIMP). In dit hoofdstuk tonen we aan dat CIMP en *CEBPA*-silenced leukemieën een aparte entiteit zijn die erg lijkt op ETP-ALL, die een overeenkomstig epigenetisch landschap bevatten in plaats van overeenkomende gen mutaties. Het methylatie patroon van deze leukemieën lijkt erg op T-ALL, maar de epigenetische en transcriptionele profielen zitten tussen AML en T-ALL in. Waarschijnlijk zijn deze afkomstig van een vroege progenitor waarbij promotor methylatie bij myeloïde TFs de myeloïde differentiatie voorkomt. Verder zorgt deze hypermethylatie ook tot verlies van CTCF binding bij CpGs, wat leidt tot veranderingen in genoom structuur met deregulatie van nabije genen tot gevolg.

Samenvattend, dit proefschrift bevestigt het belang van strikte epigenetische regulatie van hematopoëse, waarbij veranderingen in relevante mechanismen leiden tot beenmergfalen of leukemie. Met name enhancer herpositionering en afwijkende promotor methylatie leiden tot afwijkende expressie van belangrijke hematologische regulatoren zoals *CEBPA*, *EVI1* of *GATA2* in AML. Uitgebreide kennis van deze epigenetische afwijkingen en het moleculaire mechanisme hiervan, opent de deur naar de ontwikkeling van gerichte therapieën waarbij leukemogenese selectief ongedaan wordt gemaakt en de normale hematopoëse wordt hersteld.

LIST OF PUBLICATIONS

1. A. Tanaka, T. A. Nakano, M. Nomura, H. Yamazaki, J. P. Bewersdorf, **R. Mulet-Lazaro**, S. Hogg, B. Liu, A. V. Penson, A. Yokoyama, W. Zang, M. Havermans, M. Koizumi, Y. Hayashi, H. Cho, A. Kanai, S. C. Lee, M. Xiao, Y. Koike, Y. Zhang, M. Fukumoto, Y. Aoyama, T. Konuma, H. Kunimoto, T. Inaba, H. Nakajima, H. Honda, H. Kawamoto, R. Delwel, O. Abdel-Wahab, D. Inoue, Aberrant EVI1 splicing contributes to EVI1 -rearranged leukemia. *Blood* (2022), doi:10.1182/blood.2021015325.
2. R. Avellino*, **R. Mulet-Lazaro***, M. Havermans, R. Hoogenboezem, L. Smeenk, N. Salomonis, R. K. Schneider, E. Rombouts, E. Bindels, L. Grimes, R. Delwel, Induced cell-autonomous neutropenia systemically perturbs hematopoiesis in Cebpa enhancer-null mice. *Blood Adv.* **6**, 1406 (2022).
4. L. Smeenk, S. Ottema, **R. Mulet-Lazaro**, A. Ebert, M. Havermans, A. A. Varea, M. Fellner, D. Pastoors, S. van Herk, C. Erpelinck-Verschueren, T. Grob, R. M. Hoogenboezem, F. G. Kavelaars, D. R. Matson, E. H. Bresnick, E. M. Bindels, A. Kentsis, J. Zuber, R. Delwel, Selective requirement of MYB for oncogenic hyperactivation of a translocated enhancer in leukemia. *Cancer Discov.* **11**, 2868–2883 (2021).
5. S. Ottema*, **R. Mulet-Lazaro***, C. Erpelinck-Verschueren*, S. van Herk, M. Havermans, A. Arricibita Varea, M. Vermeulen, H. B. Beverloo, S. Gröschel, T. Haferlach, C. Haferlach, B. J. Wouters, E. Bindels, L. Smeenk, R. Delwel, The leukemic oncogene EVI1 hijacks a MYC super-enhancer by CTCF-facilitated loops. *Nat. Commun. 2021 121*. **12**, 1–13 (2021).
6. F. Taube*, J. A. Georgi*, M. Kramer, S. Stasik, J. M. Middeke, C. Röllig, U. Krug, A. Krämer, S. Scholl, A. Hochhaus, T. H. Brümmendorf, R. Naumann, A. Petzold, **R. Mulet-Lazaro**, P. J. M. Valk, B. Steffen, H. Einsele, M. Schaich, A. Burchert, A. Neubauer, K. Schäfer-Eckart, C. Schliemann, S. W. Krause, M. Hänel, R. Noppeney, U. Kaiser, C. Baldus, M. Kaufmann, S. Herold, F. Stölzel, K. Sockel, M. von Bonin, C. Müller-Tidow, U. Platzbecker, W. E. Berdel, H. Serve, G. Ehninger, M. Bornhäuser, J. Schetelig, C. Thiede, CEBPA Mutations in 4708 Patients with Acute Myeloid Leukemia - Differential Impact of bZIP and TAD Mutations on Outcome. *Blood* (2021), doi:10.1182/blood.2020009680.
7. **R. Mulet-Lazaro**, S. van Herk, C. Erpelinck, E. Bindels, M. A. Sanders, C. Vermeulen, I. Renkens, P. Valk, A. M. Melnick, J. de Ridder, M. Rehli, C. Gebhard, R. Delwel, B. J. Wouters, Allele-specific expression of GATA2 due to epigenetic dysregulation in CEBPA double-mutant AML. *Blood*. **138**, 160–177 (2021).
8. S. Ottema*, **R. Mulet-Lazaro***, H. B. Beverloo, C. Erpelinck, S. van Herk, R. van der Helm, M. Havermans, T. Grob, P. J. M. Valk, E. Bindels, T. Haferlach, C. Haferlach, L. Smeenk, R. Delwel, Atypical 3q26/MECOM rearrangements genocopy inv(3)/t(3;3) in acute myeloid leukemia. *Blood*. **136**, 224–234 (2020).

* These authors contributed equally to this work

ABBREVIATIONS

4C-seq	Circularized Chromosome Conformation Capture using sequencing
5mC	5-methylcytosine
AID	Activation-Induced Deaminase
ALL	Acute Lymphoblastic Leukemia
AML	Acute Myeloid Leukemia
ASE	Allele Specific Expression
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
B-ALL	B-cell ALL
BM	Bone Marrow
bp	Base pair
bZIP	Basic leucine zipper
CAR	CXCL12-abundant reticular (cells)
CEBPA DM	Double mutant CEBPA
CEBPA SM	Single mutant CEBPA
CFS	Common Fragile Sites
CFU	Colony Forming Unit
CGI	CpG Island
CHIP	Clonal Hematopoiesis of Indeterminate Potential
ChIP-seq	Chromatin Immunoprecipitation using sequencing
CIMP	CpG Island Methylator Phenotype
CLP	Common Lymphoid Progenitor
CML	Chronic Myeloid Leukemia
CMP	Common Myeloid Progenitor
CNA	Copy Number Alteration
CNG	Copy Number Gain
CNL	Copy Number Loss
CNV	Copy Number Variant
CRE	Cis-regulatory element
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CUT&RUN	Cleavage Under Targets and Release Using Nuclease
DBD	DNA-binding domain
dCas9	Dead Cas9
DCE	Downstream Core Element
DHS	DNAse I Hypersensitive Site
DI	Differential Interaction
DMR	Differentially Methylated Region
DN	Double Negative (thymocyte)

DNA	Deoxyribonucleic Acid
DP	Double Positive (thymocyte)
DPE	Downtream Promoter Element
DSB	Double Strand Break
ED	Effector domain
EGA	European Genome-phenome Archive
eRNA	Enhancer RNA
ERRBS	Enhanced RRBS
ETP-ALL	Early T-cell precursor ALL
FAB	French-American-British (classification)
FACS	Fluorescence-activated cell sorting
FDR	False Discovery Rate
FISH	Fluorescence In Situ Hybridization
FPM	Fragments Per Million
G-CSF	Granulocyte Colony Stimulating Factor
GFP	Green Fluorescent Protein
GM-CSF	Granulocyte/Monocyte Colony Stimulating Factor
GMP	Granulocyte/Monocyte Progenitor
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
GTF	General Transcription Factor
H3K27ac	Acetylation of lysine 27 in histone H3
H3K27me3	Trimethylation of lysine 27 in histone H3
H3K4me3	Trimethylation of lysine 4 in histone H3
HAT	Histone Acetyl Transferases
HDAC	Histone Deacetylases
HELP	HpaII tiny fragment Enrichment by Ligation-mediated PCR
HSC	Hematopoietic Stem Cell
HSPC	Hematopoietic Stem and Progenitor Cell
IGH	Immunoglobulin Heavy chain
Indel	Insertion/deletion
kb	Kilobase
KD	Knockdown
KDM	Lysine demethylase
KMT	Lysine methyltransferase
KO	Knockout
LMPP	Lymphoid-primed MPP
LSC	Leukemic Stem Cell
LSK	Lineage- Sca-1+ c-Kit+ (cells)

LT-HSC	Long Term Hematopoietic Stem Cell
Mb	Megabase
MBD	Methyl Binding Domain
MBP	Methyl Binding Protein
MCIp	Methyl-CpG ImmunoPrecipitation
MDS	Myelodysplastic Syndrome
MDS	Multidimensional Scaling
MeDIP	Methylated DNA ImmunoPrecipitation
MEP	Megakaryocyte-Erythroid Progenitor
MPAL	Mixed Phenotype Acute Leukemia
MPP	Multi-Potent Progenitor
NCP	Nucleosome Core Particle
ncRNA	Non-coding RNA
NFR	Nucleosome Free Region
NGS	Next Generation Sequencing
NHEJ	Non-Homologous End Joining
NK	Natural Killer
OXPHOS	Oxidative Phosphorylation
Padj	Adjusted p-value
PAMPs	Pathogen-Associated Molecular Patterns
PCA	Principal Component Analysis
PcG	Polycomb Group
PCR	Polymerase Chain Reaction
PGC	Primordial Germ Cells
PIC	Preinitiation Complex
PRC	Polycomb repressive complex
P-TEFb	Positive Transcription Elongation Factor b
PTM	Post Transcriptional Modification
PWM	Position Weight Matrix
qPCR	Quantitative PCR
RNA	Ribonucleic acid
RNA pol	RNA polymerase
RNA-seq	RNA sequencing
RPKM	Reads Per Kilobase per Million
RRBS	Reduced Representation Bisulfite Sequencing
rRNA	Ribosomal RNA
sAML	Secondary AML
SCN	Severe Congenital Neutropenia
scRNA-seq	Single Cell RNA-seq

SD	Standard Deviation
SE	Super-Enhancer
sgRNA	Single Guide RNA
shRNA	Short Hairpin RNA
SLAM	Signaling Lymphocyte Activation Molecule
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide variant
ST-HSC	Short Term Hematopoietic Stem Cell
t(3;3)/inv(3) AML	AML with t(3;3)/inv(3) rearrangements
TAD	Topologically Associating Domain
T-ALL	T-cell ALL
tAML	Therapy-related AML
TBP	TATA-box binding protein
TF	Transcription Factor
TFBS	Transcription Factor Binding Sites
TLR	Toll-Like Receptor
TMM	Trimmed Mean of the M-values
TPM	Transcripts Per Million
t-SNE	t-distributed Stochastic Neighbor Embedding
TSS	Transcriptional Start Site
UMAP	Uniform Manifold Approximation and Projection
UTR	Untranslated Region
VAF	Variant Allele Frequency
WES	Whole Exome Sequencing
WGBS	Whole Genome Bisulfite Sequencing
WGS	Whole Genome Sequencing
WHO	World Health Organization
ZF	Zinc Finger
ΔTAD	Variable TAD

PHD PORTFOLIO

DESCRIPTION	YEAR	ECs
Courses and workshops		
Galaxy for NGS	2017	1.00
Scientific Integrity	2022	0.30
Principles of Research in Medicine and Epidemiology	2022	0.70
Scientific activities at the Department of Hematology		
Presenter at Hematology department work discussions	2017-2022	2.50
Presenter at research group work discussions	2017-2022	7.50
Attendee at Department work discussion	2017-2022	5.00
Attendee at Research group work discussions	2017-2022	2.00
Department Bioinformatics/Genomics Meeting	2017-2022	5.00
Journal Club (attendance)	2017-2020	1.00
Presenter in the Journal Club of the department	2017-2020	1.50
Scientific conferences		
American Society of Hematology, San Diego (poster)	2018	0.50
American Society of Hematology, San Diego (attendance)	2018	1.20
European Hematology Association, Amsterdam (attendance)	2019	1.20
American Society of Hematology, Orlando (2x poster)	2019	0.50
American Society of Hematology, Orlando (attendance)	2019	1.20
Hematology Lectures (in-person)	2017-2020	3.00
Dutch Hematology Congress (oral presentation)	2021	0.50
Dutch Hematology Congress (attendance)	2021	0.60
European Hematology Association, virtual (oral presentation)	2021	0.50
European Hematology Association, virtual (attendance)	2021	1.50
Virtual Hematology Lectures (attendance)	2021-2022	4.00
Total EC		**41.20**

ABOUT THE AUTHOR

Roger Mulet Lázaro was born in Barcelona (Spain) on the 8th of March of 1990. He studied Biotechnology at the Universitat Autònoma de Barcelona between 2008 and 2012. During the last year of his undergraduate studies, he did an internship at Anaxomics Biotech, a Spanish start-up that uses systems biology to model human pathophysiology through the integration of protein interaction networks and manually curated information from the literature. Upon completion of the internship, having finished his undergraduate studies, he was hired with the role of junior Project Manager (2012-2015). At the same time, Roger started a Master in Drug Research and Development at the University of Barcelona (2012-2013). He graduated with a thesis entitled "Drug repositioning to treat amyotrophic lateral sclerosis using systems biology", based on the research he was conducting at Anaxomics Biotech. The following two years, he remained working full-time at Anaxomics Biotech, where his duties included management of research projects, maintenance of curated disease databases, analysis and interpretation of bioinformatics predictions, involvement in regulatory affairs and writing of scientific grants and articles. During this time, he also became intrigued by the possibilities of bioinformatics. Therefore, in 2015 Roger left Anaxomics and enrolled in a Master in Bioinformatics at Universitat Autònoma de Barcelona (2015-2016). As a part of this program, he conducted research under the supervision of Prof. Antonio Barbadilla and Dr. Sonia Casillas at the Bioinformatics of Genome Diversity group, where he studied genomic and epigenomic variation in human populations using data from the *1000 Genomes* consortium. This work resulted in a master's thesis entitled "Comparative genomics and epigenomics: Pipeline for the description of genome- and epigenome-wide patterns of variation in humans". Wishing to continue with his scientific career abroad, Roger next obtained a position as bioinformatician at the Department of Hematology of the Erasmus Medical Center, in Rotterdam (The Netherlands). There, he further developed his bioinformatics skills and became familiar with epigenetics, high throughput sequencing and leukemia. After one year, he started a PhD program in the same department under the supervision of Prof. Ruud Delwel and Dr. Bas Wouters (2017-2012). Through his doctoral research, he investigated epigenetic mechanisms that control hematopoiesis in both health and disease, with a focus on acute myeloid leukemia (AML). The program comprised three main research lines: (1) transcription factors in healthy hematopoiesis, (2) enhancer hijacking in AML, and (3) epigenetic changes in cis-regulatory elements. The results of this work are presented in this thesis.

ACKNOWLEDGEMENTS

Starting a PhD is a daunting prospect. So much to learn, so much to do; at first the end is so distant it almost seems out of reach. Yet, as is often the case in life, when the journey is over one looks back and marvels at how fast time has gone by. And with this retrospective view comes the realization that the whole enterprise would not have been possible without the assistance of others. This assistance comes in many forms: mentorship, intellectual assistance, teaching, emotional support and, of course, occasional distraction and fun. Therefore, I want to thank here everyone who has helped me along the way – my colleagues, my friends, and my family. I am lucky to have you all.

I would like to start by thanking my promotor, **Ruud Delwel.** Dear Ruud, it has been a true pleasure to work with you all these years. I ended up as a PhD candidate in your group almost by accident. I was originally hired as a bioinformatician for the department, but I tended to focus on epigenetics data for the Delwel group. And when the time came to consider a PhD trajectory, it simply made sense for me to join you. Thanks for adopting me. As a young researcher freshly arrived from Spain, I initially felt a bit intimidated by (multiple) award-winning Prof. Dr. Delwel. But like everyone who has met you knows, I quickly realized you are an approachable, down-to-earth and caring person. No matter my concerns, scientific or otherwise, I have felt at ease discussing them with you. There were times when your other responsibilities kept you away, but even then you tried your best to be there. You even joined a few online meetings at 4 or 5 AM while traveling in the US! And during the COVID lockdown, you would regularly contact people in your team to make sure everyone was doing fine. All these examples are proof of your deep commitment to the people under your supervision. These human qualities are matched by your qualities as a scientist. I admire your intellectual curiosity and your creativity, which together result in new ideas and new angles to tackle research questions. Your tenacity, optimism and ability to detect promising discoveries inspire others not to give up and pursue new research avenues. Both your human and scientific side converge on fruitful collaboration networks with other high-caliber scientists. The whole department has greatly benefited from initiatives such as the Hematology Lectures, which have been a huge success in a virtual format as well. For all these reasons and many more, I feel privileged for having you as a mentor.

I would also like to thank my co-promotor, **Bas Wouters**. First, for putting together an exciting research program that constituted the foundation of my PhD trajectory. Second, for your valuable supervision and advice during all these years. My work has greatly benefited from your ability to focus on the important questions, which counterbalances my tendency to get lost in too many possibilities. Your medical background has also helped me judge things from angles I would not have considered otherwise. I realize it must not have been easy to reconcile my supervision with your demanding clinical duties and your involvement in clinical trials, and yet you always made time for our regular meetings. Thank you. Last but

not least, I have genuinely enjoyed working with you at the personal level. You are a kind man with a fantastic sense of humor, always quick to smile when someone (often yourself) in the room makes a joke.

I also want to express my deepest gratitude to **Ivo Touw**, who was my supervisor during my first year as a bioinformatician in the department. Together with Ruud and Mathijs, you are the reason I got the job, and eventually started my PhD a year later. Since our first meeting, and throughout all these years, I have been consistently impressed by your deep knowledge of hematology, and more broadly, biomedical sciences. Furthermore, you are a critical thinker with an uncanny ability to identify potential weaknesses in scientific research, making your feedback extremely valuable. A feedback that you are, in fact, very willing to share -- without mincing words. Whenever I had to present my work in the department, I prepared myself for your hard questions, a feeling that I am sure is shared by many others. However, even if it is harsh at times, your criticism is always valid. It is for that reason that I showed you my abstract for ASH2019, which you essentially dismissed as unacceptable. It certainly improved thanks to your remarks, and it is possibly for this reason that it was finally accepted. Now you have also been part of my Doctoral Subcommittee, and once again I have to thank you for accepting this responsibility and your valuable suggestions.

I owe a special debt of gratitude to **Mathijs Sanders**, thanks to whom I joined the department in the first place. Thanks for placing your trust in me during your stay in the Sanger Institute. As a freshly graduate from a master in bioinformatics coming all the way from Barcelona, I had little experience with the analysis of next generation sequencing data, even less so with leukemia. My research had been focused on population genetics and I had only worked with processed data, after all. However, you taught me the ropes and provided me reading material that would prove extremely helpful. I have to confess I never got very far with your Linear Algebra book, but it is definitely something I want to pick up again someday. You are a brilliant scientist with extensive, multidisciplinary knowledge, encompassing mathematics, programming and biology. One of your biggest assets is your passion for science and a never-ending curiosity to understand how things work. Your insightful questions and comments during work discussions have been very valuable, helping me reconsider the results of my analyses under a different light. I am sure you will be a great mentor to your current and future PhD candidates under your supervision. But I cannot finish without also thanking you for the fun during our “bioinformatics” lunches, in which your (often weird) personal stories never ceased to amuse me.

Dear **Remco**, old chap, your contribution to this thesis has been pivotal. How could I have made it without the HP computer with Ubuntu Mate I got from you? Or without the many software tools you wrote, including extract_vaf_matrix_roger or annotate_bam_statistics_roger. I started as a bit of a noob, but you have taught me lessons I will never forget. At the professional level, I have learnt about computer hardware, network infrastructure and programming. At the personal level, my Dutch has been enriched with terms like *flapdrol*,

gekbek, pannenkoek or *apekop*. Jokes aside, you are a fantastic professional with a deep understanding of both hardware and software, who works hard to ensure that all researchers in the department are "satisfied customers". The server infrastructure you designed has made it possible for me to analyze and store TBs of data. Perhaps more importantly, you have been an amazing colleague and a true friend. I thoroughly enjoyed our endless discussions about nuclear power plants, public transportation, space travel, electric cars, healthy food, PC components and many other topics. Your pranks never failed to put a smile on my face, even the time you made me think you had stolen one of my (limited edition) Twix bars. In retrospect, maybe I could have finished my PhD a bit earlier if not for the constant distraction ;). But it has been worth it. Thanks for the countless good times in all these years together, in the same small office. Finally, I have to admit you were right: SSDs are still more expensive than spinning hard drives.

To the rest of bioinformaticians in the department – **Gregory, Chiel, Jolinda**. We have not spent that much time together, especially because you started in the middle of a lockdown. However, it has been enough for me to realize you are all brilliant, smart and meticulous people who bring unique sets of skills and personalities to the team. Chiel, you deserve a special mention for patiently listening when I interrupt your work to talk about any random topic. **Elodie**, you will probably be in Norway by the time this words are printed, but thanks to you as well for making dramatic improvements on the organization of the server and building fantastic websites for both diagnostics and research. I wish you a very successful PhD in the North, if you survive the cold.

Dear **Leonie**, we started working at the department almost at the same time, and during these years we have collaborated on a number of projects. I still remember how we struggled with the mess of 3q-capture and 4C-seq data for 3q26-rearranged AML patients! I still find myself using your summary table when I have to deal with those identifiers… I admire your dedication to science, in particular your constant efforts to stay up to date with the latest developments. And you have helped me stay up to date as well by regularly sharing research articles and reviews about epigenetic regulation. Thanks a lot for that! Besides, you are a major driving force in the department and the Delwel group, taking care of organizing the Friday meetings and other activities. I appreciate your constant efforts to make sure everyone remembers about those and feels included in the team. In fact, as I write this, I have just received an email from you about the upcoming DHC. I wish you best of luck with your grant on *EVI1* regulation in ovarian cancer and, more generally, with your future scientific career. I hope we get to collaborate in the future again.

Dear **Stanley**, you are an excellent professional – smart, precise, rigorous and hard-working. Proof of that is the top-notch data you have generated all these years, not only for the Delwel group, but for others as well. In fact, my own research would not have been possible without those data. Thank you. You have also proven to be very skilled at implementing new techniques and developing new models, such as Cut&Run, the DNMT3A-

dCas9 system, the MUTZ3-EVI1-GFP/GATA2-mCherry cell line, and many more. It has not always been easy, but no matter the obstacles, you always try your best, doggedly keeping at it until you find a solution or you determine it is not feasible. I suspect you have not given up on Cut&Run just yet ;)

Dear **Sophie**, I finally followed your advice: I put together a bunch of papers, wrote a couple of extra chapters and, voilà, I had a thesis. Easy! Of course, some of those papers are the result of your own hard work. You did an amazing job making sense of atypical 3q26 AMLs and identifying key common features shared with classical 3q26 cases, a nightmarish task considering how heterogeneous they are and their ridiculously complex rearrangements. Your work on the generation of K562 cells with a t(3;8) model was superb, and it will probably be an inspiration for similar models in the future. And all along you managed to reconcile your professional activities with your personal life. You are proof that it is possible to do great research without renouncing to a good work-life balance. It was a pleasure to collaborate with you, also at the personal level. You are a really fun person and I admire your no-nonsense and straightforward attitude, while at the same time being easy-going and approachable. Best of luck with all your current and future endeavors.

Dear **Roberto**, you have been an inspiration since I first joined the department. I admire your devotion to science, your hard work and your fascinating ability to remember complete literature references (including authors, year and journal!). It is always instructive to talk to you. You have played a prominent role in the work leading to this thesis. First, you generously invited me to share authorship of the *Cebpa* mouse model paper, even though I had only joined later. Second, you were always available to discuss my ongoing projects, give me advice and critically read my manuscripts. Importantly, you are a caring and empathetic person, always willing to listen and give support. Altogether, it has been a privilege to work with you, and I am glad we have kept in touch after your departure. I am looking forward to reading the results of your present work.

I am also indebted to the rest of the current and past members of the Delwel group, a veritable “dream team” of talented people. **Claudia Erpelinck**, your ability with ChIP-seq, 4C and other chromatin techniques is unparalleled. You are also extraordinarily kind and cheerful, always quick to smile. Thanks for inviting all of us to your home for a Christmas dinner, it was my first (and so far only) time having winter BBQ! **Dorien**, you are a passionate scientist with a knack for both wet lab and bioinformatics, a rare combination of skills that makes you very valuable in any team. And you are a fantastic baker to boot, as I have only recently discovered (though the *dulce de leche* should have been darker!). Best wishes on the rest of your PhD and future scientific career. **Marije**, your expertise and dedication have been essential for multiple research programs presented in this thesis, including the 3q26 AML work and especially the *Cebpa* mouse studies. Throughout my PhD, several people have also joined the Ruud’s group; sometimes only briefly, but making important contributions nonetheless. Among them, **Andrea** deserves a special mention for her indispensable

experiments in the 3q26 AML studies. *Mucha suerte con todo!* Finally, **Maikel**, you did a wonderful job analyzing RNA-seq and 3q-seq data, demonstrating not only technical skill, but also creativity and critical thinking. You have a great career ahead of you.

Dear **Eric**, you are one of the cornerstones of the department. Your role as operator of the Illumina machines (the Hiseq when I first arrived, now the Novaseq) is indispensable to many, including myself. As a matter of fact, all the chapters in this thesis rely heavily on NGS data generated in the department. Thank you for making it possible. Thank you as well for showing me how the Hiseq 2500 worked when I took you up on the offer you made on the first day we met. In addition to your direct involvement in the sequencing, you also make sure the quality controls are in order and take an interest in the downstream analysis, demonstrating familiarity with a surprising number of projects and research lines. I also deeply respect how your inquisitive mind keeps up to date with the latest technological developments and scientific discoveries. At the personal level, I would like to express my gratitude for the counsel and advice you have given me, especially in my starting years as a PhD. And for your (normally fruity) cakes, of course!

Dear **Peter**, congratulations on the amazing job you do managing the diagnostics and research arms of your group, both of which achieve excellence in their respective goals. The exhaustive analyses and careful organization carried out by your team are an enormous asset to the whole department, and have definitely been critical in my own projects. Although it has been a few months since then, I would also like to reiterate my felicitations on your appointment as association professor, which is very much deserved. Finally, I am also grateful for your valuable feedback during the Delwel-Valk meetings and Friday work discussions. **Emma Boertjes,** I am impressed by the number of projects you can handle at the same time, covering diverse topics such as splicing, TP53 mutations and 3q26-rearrangements. Good luck! **Francois**, your painstaking job analyzing the results of the diagnostics pipeline is extremely valuable for researchers and patients. Thank you for providing data when I needed it. **Tim**, kudos for a PhD trajectory full of successes, including a NEJM and a Blood paper; you will undoubtedly have a wonderful thesis and a very promising career. To the rest of the current and present members in the Valk group, thanks for your inestimable hard work.

In addition to the aforementioned, the department of Hematology of Erasmus MC counts with an outstanding rooster of principal investigators: **Ruben, Emma de Pater, Tom, Marc, Mojca, Frank, Moniek, Jan, Pieter** and **Bob**. Thank you for keeping a remarkable level of excellence in this department and your insightful input during work discussions. Thank you as well for sharing your knowledge in the "Back 2 Basics" series this year. I would also like to acknowledge the important role of postdocs and other senior researchers in spearheading new projects and providing guidance to starting scientists: **Mark van Duin**, **Iris**, **Lanpeng**, **Diane**. To the current PhD candidates, veterans and newcomers: **Eline**, **Madelon**, **Jacqueline**, **Maurice**, **Paola**, **Sanne**, **Sjoerd**, **Christian**, **Martijn**, **Cathelijne**, **Sabrin**,

Lorenzo, **Calvin**, **Isabel**, **Bas Laan**, **Sophie Hordijk**, **Ryanne**, **Niek**, **Chantal**, **Aarazo** and **Yujie**. Best wishes on the rest of your trajectory and your future scientific career. Even if it seems far at times, there is always light at the end of the tunnel. Finally, I would like to thank the excellent technicians of the department: **Onno**, **Dennis**, **Michael**, **Jasper**, **Melissa**, **Claire**, **Petra**, **Elwin**, **Hans**, **Natalie**, **Wendy**, **Mariëtte**. I hope I did not forget anyone, I am afraid I have not interacted much with some of you during the almost two years I spent working mostly from home.

I would also like to thank **Egied** for his outstanding job in the preparation of the layout of this thesis, as well as **Tessa** and **Leenke** for their help with administrative duties.

Furthermore, I would like to mention a few people who are no longer in the department. One of my first projects as a bionformatician in the department was the analysis of acute lymphoblastic leukemia sequencing data for **Aniko Szabo**. Thank you for the learning experience. I am also grateful to you for generously sharing the RNA-seq data from T-ALL patients for the CIMP leukemia project. **Davine**, your late afternoon visits to our office were always a welcome distraction (a *fika* of sorts, a word that in fact I learnt from you). Congratulations on your recent publication in JCO! You deserve it after a long and arduous journey. **Ping**, I thoroughly enjoyed our conversations on various topics, including history and, of course, food. Thanks for your valuable advice on post-PhD options. And once again, I was very impressed by the beautiful design of your thesis. **Mira**, you are a delightful person, I still remember how you brought Christmas cards for everyone in the department. **Emanuel**, **Keane** and **Cansu**, you did a fantastic job in your PhD, best of luck with your future scientific careers.

Outside the department, I would like to thank **Claudia Gebhard**. You have played an essential role in several chapters of this thesis, which are fruit of the collaboration between our groups. In fact, the first project in my PhD was the study of the enhancer landscape of AML using H3K27ac ChIP-seq data generated at your lab. During all these years I have been consistently impressed by your hard work, diligence and attention to detail. However, sometimes you should take things a bit easier. Most things turn out just fine, no need to panic! Somewhat ironically, we had already been working together for years when we finally met at ASH 2018 in San Diego, on the other side of the world. We have talked about it so many times, but we haven't found the occasion to meet in Regensburg or in Rotterdam. I'm still waiting for your invitation! Also in Regensburg, I would like to thank **Michael Rehli** for starting some of the projects I was involved in, as well as for key insight and advice. It was a privilege to work with you. **Alexander**, you have done an amazing job analyzing and integrating multi-omics data to study the effects of cohesin alterations in AML. To the rest of the team, including **Nick** and **Ute**, thanks for your hard work.

I would also like to thank a number of mentors I have had during my formative years. **Antonio Barbadilla** and **Sonia Casillas**, thank you for accepting me as an intern in your research group. Not only did I learn a lot about population genetics, evolution and

bioinformatics, but I also became aware of the importance of evolutionary thinking. It has certainly influenced my outlook on genetics and biology as a whole, for the better. My 11th proposition attests to it. It has been several years (time flies!), but I remember with fondness our discussions and journal clubs. Antonio, I would also like to thank you for creating and coordinating the MSc in Bioinformatics, which allowed me to transition into this field. Dear **Dolores**, thanks for instilling in me the passion for science that put me on this path. It was your amazing teaching and your encouragement that convinced me to choose a bachelor´s degree in Biotechnology. As a young high student, I conducted my first (admittedly amateur) research project under your supervision, involving Drosophila subjected to caloric restriction to evaluate their survival in relation to controls. I remember vividly those evenings in the school lab, putting flies to sleep with ether to count them. In retrospect, that was a foundational moment in my career as a scientist. Little did I know that I would end up pursuing a PhD in the Netherlands.

Last, but not least, I want to thank my family and friends for their unwavering support. To my dear parents, **Isabel** and **Ricard**, thank you for your love and unconditional support. As teachers yourselves, you taught me the importance of education and studying early on in life, which has undoubtedly helped me get to this point. I realize it must have been hard for you to see me leave and to stay apart all these years. I have missed you as well. To my sister, **Laia**, I miss you too, even if I don't always show it. I am glad you have found something you like and decided to pursue it despite the obstacles. Even though we have not met very often, especially during the COVID19 lockdown, I also carry my extended family in my heart – my cousins **Laura**, **Aina**, **Nil**, **Júlia**, **Pol**, **Joan**; my uncles and aunts **Gemma**, **Josep**, **Alícia**, **Miquel**, **Joan**, **Natàlia**. And very especially my grandma **Roser**, who always makes sure I have a good supply of *ibérico* ham when I travel back. And of course those who are not there anymore, but are deeply missed: my grandparents **Damiana**, **Vidal** and **Joan**.

Dear **Xi**, you have been an essential pillar of my life throughout this entire journey – a close friend, a confidant, a role model and much more. In many ways I am here because of you. I have enjoyed our culinary adventures, the small yet meaningful daily conversations, our trips together. I treasure those precious memories. Thank you for your company, your support, your trust and uncountable good moments. Thank you for being there no matter what. You are my person, you will always be my person. Dear **Juncai**, my favorite opera singer, what a rollercoaster it has been since our lives first crossed. Who could have predicted we would end up being so close? We have not always seen eye to eye, we have even grown apart at times, but we always find the way back. I will be forever grateful to you for helping me overcome a tough period in my life. Thank you for your affection and patience, I love you. Dear **Kiki**, you will not be reading this -- you are a cat, after all! But you have made my days working from home much easier, providing a much welcome distraction when I hit a stumbling block or I simply needed minutes away from my screen.

I am also indebted to the rest of my friends in the Netherlands, all of whom have left

their unique footprint on my life. **Daniel**, I have enjoyed our recurrent debates about all kinds of topics, your well-though arguments never fail to give me pause for thought. Also, I am still amazed by how much you can eat! **Wilson**, I thought you were going to finish your PhD first, but I beat you to the punch! ;) Funny that you ended up working at EHA, I hope we meet at the next conference. **Yuen**, thanks for being a good friend, always up for a chat or a walk in the park. Best of luck with your PhD as well, you will crush it. **Leo**, we have had our disagreements, but I appreciate the good times we shared.

Despite the distance, I count myself privileged for keeping in touch with my friends from Spain. **Xavi**, our weekly gaming sessions have become a tradition going on for already more than 5 years. I must confess I had my doubts when you had a baby, but I appreciate you manage to find time even with a toddler in the house. Also, I appreciate your sense of humor, you always manage to put a smile on my face even with the same old jokes. **Edu**, you are the one person with whom I feel free to discuss absolutely everything. I appreciate your open-mindedness and critical thinking, as well as your willingness to recognize and minimize preconceived notions. Moreover, I admire your dedication to any personal and professional projects. Thank you for always finding time to meet whenever I am in Barcelona! **Arnau**, you are my oldest friend, we have known each other almost for as long as I can remember. In school, we were in the same class since we were 6 until we turned 18. We were thick as thieves back then, to the point some people mixed us up! I still smile with nostalgia when I recall our many shared memories, such as the class-wide role playing we called “el joc de la classes” or the time you ordered a steak in our trip to Italy, while the rest of us survived on humble sandwiches. Although we have both changed over the years, I am glad we remain good friends – it’s no small feat! **Héctor**, I knew we would get along since the first time we met in college. We had so much in common! No matter how much we talked, I always had the feeling there was more to say. Of course, our classmates may not have been so happy about it… Although our scientific career has taken us to different parts of the globe, the rare occasions in which we meet bring me back to those old days when we could talk nonstop for hours. Funny enough, we both ended up switching to Bioinformatics, and recently you have also been working on leukemia. Who knows what other coincidences await in the future.

These acknowledgements inevitable fall short to recognize the influence every single person has had in my life. Relatives, friends, teachers, colleagues, even short-lived acquaintances… The ways in which human interactions change our lives forever are many and impossible to capture in a few pages. However, I want to express my gratitude once again to all of you. Thank you.


GTGAAGACGTGTATCGACCA
CGGCTGAGATCCGCAACGTAACAAGCGCTGTGTAAGGTCGG