WWU MÜNSTER

WESTFÄLISCHE WILHELMS-UNIVERSITÄT
INSTITUT FÜR MOLEKULARE TUMORBIOLOGIE

MATTHIAS LUTZ ZEPPER

Epigenetic characterization of murine Dnmt1-deficient MLL-AF9 leukemia

DNA methylation in large regions and at cis-regulatory elements dissected in the Dnmt1$^{-/chip}$ mouse model

2020

Universität Münster BIOLOGIE

Supplementary methods, figures and results

BIOLOGIE

Epigenetic characterization of murine Dnmt1-deficient MLL-AF9 leukemia

DNA methylation in large regions and at cis-regulatory elements dissected in the Dnmt1$^{-/chip}$ mouse model

---

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften im Fachbereich Biologie der
Mathematisch-Naturwissenschaftlichen Fakultät der
Westfälischen Wilhelms-Universität Münster

vorgelegt von

MATTHIAS LUTZ ZEPPER
aus **LUDWIGSBURG**

2020

Dekanin:	Prof. Dr.rer.nat. Susanne Fetzner
Erster Gutachter:	Prof. Dr.rer.nat. Frank Rosenbauer
Zweiter Gutachter:	Prof. Dr.rer.nat. Joachim Kurtz
Tag der mündlichen Prüfung:	12. März 2020
Tag der Promotion:	. . . . .

Contents

Supplementary Chapter 1

Introduction to enhancers and other cis-regulatory elements

1.1 Gene-regulatory genomics

Protein-coding genes occupy only marginal fractions of the total sequence length in most eukaryotic genomes [1]: For those of mouse and human it is within 2 % to 3 %. Scientists initially demeaned the remainder as "junk DNA" due to larger stretches of repetitive sequence and parts of viral origin bearing witness to the developmental history [2, 3]. From an evolutionary perspective, most species have favored genome integrity over active reduction of surplus sequence, unless nutrient poor conditions force to economize on phosphorous, nitrogen and carbon during DNA replication. The latter is exemplified by the genome of the carnivorous aquatic plant *Utricularia gibba*, whose 28 000 genes account for 97 % of its just 82 Mbp (mega-base pairs) genome size [4,5]. The human genome is 40 x larger (3200 Mbp), but comprises just moderately more genes (46 500). A mouse genome is highly similar in size and gene content, but both are easily surpassed by plant and amphibian genomes, which may reach sizes of 150 Gbp or more [6].

This highly variable content of non-protein-coding DNA sparked a still ongoing lively debate over the amount of functional DNA in our genome [7–9]. The better part of mouse and human genomes is to a variable degree transcribed into RNA [10,11], which however does not necessarily imply function [12]. As it has been estimated that just 8.2 % of the human genome is presently subject to negative selection [13], thus faithfully preserved throughout evolutionary history, the truly functional share might be in the single digit range.

In any case the functional fraction comprises the protein-coding genes, various operational RNA molecules (rRNA, tRNA, snRNA, piRNA, lincRNA and others), rather structural parts like telomeres and centromeres and also non-coding parts of DNA that are "regulatory" [14]. In the broadest sense, the regulatory parts switch genes on and off or modulate the amount of transcript or protein output and are vital for the establishment of tissue- and/or cell-specific expression. Some regulation occurs after transcription took place, like that by microRNAs (miRNA), but a lot of control is directly exerted on the

stage of messengerRNA (mRNA) production. Most of these mechanisms require the presence of regulatory sequences in the vicinity of the gene to be addressed. Such sequences are termed *cis-regulatory elements*, because they are located on the same DNA molecule before, within or after the gene itself. To act on the gene, they in general make contact with its promoter resulting in a DNA loop [15, 16]. Typically one gene is targeted by several cis-regulatory elements and one element may also be involved in the regulation of different genes [17–19].

1.1.1 Classes of cis-regulatory elements

The repertoire of cis-regulatory elements is quite diverse and three major categories, namely enhancer, silencer and insulator have been distinguished classically. The locus control regions are sometimes considered to be another separate category, however they altogether may be revised in future, since new findings challenge the classical schema: Next-generation sequencing based methods allowed to scrutinize thousands of regulatory elements in parallel and in many different cell types and developmental states. These investigations casted doubts on the concept that a particular element inherently functions in one determined way. Instead they supported the notion that cis-regulatory elements possess a regulative potential (which may vary between individual elements) and the epigenetic state, chromatin conformation and expressed transcription factors ultimately decide the outcome [20].

Enhancers: A regulatory sequence, which will orientation-independently increase the transcription rate of nearby genes when active [21, 22]. It recruits positive transcription activators, proteins with the capability to support the production of the mRNA transcript [↪ section 1.2].

Silencers: A sequence-specific element that exerts a negative effect on the transcriptional output of the targeted gene. By recruitment of repressive factors (like the KRAB zinc finger proteins [23, 24]), silencers interfere with the correct assembly of the transcription preinitiation complex (PIC) at the gene promoter [25], a basal protein complex that is required to transcribe DNA sequence into RNA. Silencing can also be achieved by blockage of RNA elongation, as in the osteocalcin gene, where the relative abundance of two competitively binding proteins decides the outcome [26,27]. Particular CpG islands may form another group of negative regulatory sequences and serve as polycomb response elements (PREs), although the existence of PREs in mammals (contrary to *Drosophila*, where they are well characterized) remains debated [28–30]. Silencers, like enhancers, act in an orientation-, position-, and distance-independent manner.

Insulators: DNA sequence elements that shield a gene's promoter from the ramifications of nearby cis-regulatory elements or serve as heterochromatin boundary. Adjacent repressive chromatin tends to propagate into a transcriptionally active euchromatic region unless an insulator element protects it [31]. The active chromatin configuration is maintained by recruitment of various insulating proteins like

USF1, VEZF1 and potentially CTCF [32–34]. Furthermore insulators can exhibit a cis-element-blocking function, when they are located in between a promoter and an enhancer or silencer. The chicken β-globin insulator (cHS4) is a well studied vertebrate insulator, which possesses both heterochromatic barrier, silencer- and enhancer-blocking capabilities [35–37]. The advent of next-generation sequencing based methods to quantify incidences of chromosomal contacts (particularly Hi-C [38]) has introduced the concept of topologically associating domains (TADs) [39, 40] [reviewed in 41]. Chromosomal contacts are frequent within a TAD, but rarely occur outside of it. Nevertheless not all TAD boundaries qualify as classical insulators, as they can for example also occur at housekeeping gene promoters [42, 43]. Furthermore Hi-C analyses after CTCF knockdown have shown that CTCF mediates transcriptional insulator function through enhancer blocking, is implicated in TAD organization and chromatin folding but not relevant to rein heterochromatin spreading [44]. Thus, shielding promoters from nearby cis-regulatory elements and limiting heterochromatin spreading might be independent functions.

Locus control region: It opens condensed heterochromatic domains and primes whole gene clusters for activation. Once a LCR has induced conformational change of chromatin, local enhancers within the cluster sustain and fine-tune the transcription of the genes. Locus control regions act over large distances as long-term on-switches in the chromosome, since the reestablishment of the repressed chromatin state requires the cell to pass through S-phase [45]. In experiments, LCRs, which have been artificially cloned outside of their native context, subsume properties of both transcriptional enhancers and chromatin insulators. It is however disputed, whether they exert all of those functions in a regular chromosomal environment and if they should merely count as strong enhancers [46]. A well-studied LCR activates the β-globin locus when erythroid lineage commitment takes place [19, 47, 48]. Newer publications coined the terms *super enhancer* [49] or *stretch enhancer*, which were also applied on known LCR regions. Thus, both names are mostly synonymous, although LCRs may be a more specific subset.

Other: On top of that there are various further motifs and genomic segments , which can justifiably considered to be potential separate entities of cis-regulatory or cis-acting elements. This includes response elements to a variety of factors like hormones, cytokines (interferones, interleukins), hypoxia-inducible factors, retinoic acid or vitamin D. Furthermore upstream open reading frames (uORFs) and upstream AUG (uAUGs) located in the 5'UTR of mRNAs [50, 51] as well as elements for splicing control [52] shall be mentioned briefly.

1.2 Characteristics of eukaryotic enhancers

Operationally, enhancers augment the activity of a nearby promoter and are enriched in recognition motifs for sequence-specific transcription factors. Since the orientation and

to a large extent also the exact position relative to the promoter is insignificant, they are considered to be orientation- and position-independent. Anyhow, most enhancers are thought to reside in the vicinity of the targeted promoter, either a few thousand bases upstream or (often in the first two introns of the gene) downstream. However, also long-range promoter contacts are common [53,54], as illustrated by the regulatory framework of the MYC locus [55] or the SHH gene [56].

1.2.1 Mode of action

The mechanism of enhancer action requires a change of the three-dimensional chromatin structure and the formation of a DNA loop. In doing so, enhancers increase the concentration of transcriptional activators near the promoter, but the temporal progression of the process is poorly understood. The discovery of preexisting chromatin looping, which precedes the actual signaling [57] has challenged the previous model [58] that solely specific transcription factors govern the looping [59, 60]. Furthermore formation and dissociation of loops must undergo tight sequential coordination as one gene is typically targeted by several cis-regulatory elements and one element may also be involved in the regulation of different genes [17,18,61]. Embryonic development [62] and hematopoietic lineage commitment [63] exemplify the latter.

Once enhancer and promoter have converged, coactivators facilitate proteinprotein interactions [64], modulate the activity of transcription factors, alter epigenetic marks on histones (which aid the decompaction of the chromatin fiber) [65, 66] or recruit kinases to phosphorylate the c-terminal tail of RNA polymerase II [67]. Out of the hundreds of different coactivators in the mouse and human genome, more than fifty operate on any given transcript start site (TSS) in conjunction. The centerpiece of the coactivator structure is the mediator complex [68], which is conserved from yeast to human [69,70]. For its identification and characterization, the Nobel Prize in Chemistry was awarded to Roger D. Kornberg in 2006.

Bringing enhancer-bound protein factors close to the promoter-bound preinitiation complex (PIC) is the best characterized function of enhancers, but not the sole. Besides proteins, also regulatory RNAs take part in the molecular machinery of eukaryotic transcription, some of which are linked to enhancers. Pol II-transcribed enhancer RNAs (eRNAs) are the most prominent example [↪ subsection 1.2.3], but also the Pol III-transcribed non-coding 7SK small nuclear RNA (7SK) can be recruited by enhancers [71]. 7SK has a scaffold function for an inhibitory 7SK small nuclear ribonucleoprotein (7SK snRNP) complex, which in its canonical form suppresses P-TEFb-mediated release of RNA polymerase II pausing at promoters [72,73]. At enhancers an alternate assembly of 7SK-complex dominates, which incorporates members of the BAF chromatin-remodeling complex and inhibits enhancer RNA transcription by modulating nucleosome position [74]. As the two 7SK complexes are mutually exclusive [71], it is tempting to speculate that the enhancer-bound 7SKBAF snRNP repels the canonical 7SK snRNP and permits release of promoter-proximal pausing. Indeed, the existence of anti-pause enhancers has

been shown [75], however the predominant mechanisms seem to be BRD4 and JMJD6 recruitment for destabilization of 7SK by decapping [75, 76] and acting as decoy for NELF [77].

It should be emphasized that the enhancer characteristics outlined in this section are largely limited to the genomes of higher eukaryotes. While sea urchins, which developed about 500 million years ago, employ enhancers [78, 79] and insulators [80] for transcriptional regulation, lower eukaryotes such as yeast rely on upstream activator sequences (UAS) [81–83]. They are somewhat analogous to enhancers of higher eukaryotes, but differ in their inability to activate a promoter from downstream positions. They do however function in either orientation and at variable distances from the promoter. In bacteria, a distinct form of RNA polymerase containing sigma factor 54 (σ^{54}) recognizes simple enhancers upstream of a promoter [84]. Viral genomes like those of the simian virus 40 (SV40) may contain own enhancers, which will become functional after integration into the host genome [85–87].

1.2.2 Enhancer states and activation

Before next-generation sequencing based methods were developed to identify potential enhancers genome-wide [↪ section 1.3], artificial expression systems were used to validate the ability of a DNA sequence to enhance transcription. The elements to be tested were cloned adjacent to a weak promoter, which typically originated from a different gene than the one initially associated with the regulatory region [88, 89]. When the transcription rate was quantified as readout in vitro or in vivo, even such heterologous transcripts would often recapitulate known tissue-specific regulation. Thus, selective activation was mostly imposed by the enhancer [90, 91].

The core promoter seemed to define temporal windows with the opportunity for activation, subsequently precise expression timing was imposed by enhancers [92]. This finding raised the question how enhancers are regulated. Today it is widely accepted that the correct subset is selected from the repertoire of potential enhancers by means of higher-order chromatin organization and specific pioneering transcription factors, which open up condensed chromatin and expose the enhancer to the binding of additional factors. Once operable, the formation of chromatin loops, changes in methylation and the recruitment of further transcriptional activators constitute further regulatory layers. Four enhancer states can be distinguished according to Heinz and colleagues [93]:

Closed: The cis-regulatory sequence is situated in highly compacted chromatin and is inaccessible for transcription factor binding. Furthermore it lacks characteristic histone modifications other than those associated with heterochromatin. Enhancers, which are closed in stem cells and only appear in differentiated cells are referred to as latent enhancers.

Primed: Pioneering transcription factors have facilitated the loosening of the compacted chromatin [94] and a nucleosome-free region of accessible open DNA was established. This enables the binding of cooperating sequence-specific transcription

factors and the recruitment of coactivators. Nevertheless essential cues for activation are missing.

Poised: This state is in many regards comparable to a primed enhancer, but nearby repressive epigenetic chromatin marks like H3K27me3 quiesce the site additionally. Since they enable rapid activation of a particular developmental program, poised enhancers are often found in embryonic stem cells and to a lesser extent also in tissue-specific stem cells.

Active: The core of an activated enhancer is nucleosome-free, open chromatin permits the binding of transcription factors and coactivators. Flanking nucleosomes are typically marked by H3K27ac [95] and eRNAs are produced at active enhancers.

The signature of histone marks [96, 97] and bound coactivators as well as the enhancer's propensity to generate eRNA transcripts is highly relevant. Since some of the next-generation sequencing based methods [↪ section 1.3] rely on these patterns to identify enhancers genome-wide, sensitivity and specificity of the respective method will vary and sometimes confine itself to enhancers in a particular state.

1.2.3 Enhancer RNAs

In retrospect, Northern Blot experiments of the β-globin locus control region showed transcription of enhancer elements in 1990 already, however the results were mistaken for run-through main transcript and attributed defective polyadenylation [98]. Two decades later, a study in neuronal cells uncovered bidirectional transcripts initiated at known enhancers with occupancy of co-activators and termed them enhancer RNA [99]. Around the same time another group confirmed these observations in macrophages [100]. The production of eRNAs was evidently restricted to active enhancers, since the transcript start sites were reported to be enriched in H3K4me1 and H3K27ac, but devoid of H3K27me3. If those are unavailable, H3K8ac, H3K9ac and the elongation associated H3K36me3 & H3K79me2 can serve as surrogate markers for eRNA production [101].

Enhancer RNAs were soon shown to be capped on the 5'end, short (<1 kb), bidirectional, unspliced and rapidly degraded by the exosome [102–104], which contested a relevant functional role of the pervasive transcription initiating from enhancers. Contrarily, eRNAs have been demonstrated to stabilize enhancerpromoter association and chromatin loops at steroid hormone response genes [105, 106], to be of importance for H3K4me1 and H3K4me2 deposition by MLL3 and MLL4 at *de novo* enhancers [107] and to be subject to functional methylation [108]. Thus, they exert meaningful roles at least for a subset of enhancers [reviewed in 109] and their expression generally correlates with the that of their target genes [110].

Especially the relationship and interplay of enhancer RNAs (eRNAs) and long-noncoding RNAs (lncRNAs) challenged the distinctness of enhancers and promoters and has been subject to an ongoing scientific dispute [111]. While the ability of certain lncRNAs to enhance the expression of nearby genes [112] by preventing DNA methylation [113] and

maintaining active chromatin states [114] was known, the finding that the majority of them originated from enhancer-like elements came by surprise [115, reviewed in 116]. Furthermore Kdm5c was shown to regulate whether individual elements rather exhibited enhancer-like or promoter-like activities [117]. Previously the lack of splice donors [118] proximal to enhancers was believed to preclude productive elongation of eRNAs [104]. Instead, it was shown that the adapter protein WDR82 and its associated complexes actively terminate the transcription. After knock-down of complex proteins, many eRNAs were actively elongated into lncRNAs [119]. This finding also provides a rationale for the presence of the 7SKBAF snRNP [↪ subsection 1.2.1] at enhancers. Convergent transcription, which would inevitably occur upon elongation of eRNAs at clustered elements within super enhancers, needs to be suppressed as it would otherwise trigger strong DNA-damage signaling [74, 120].

In summary, broad similarities between enhancers and promoters exist: Both may initiate directed or undirected transcription. Interacting regulatory complexes as well as the surrounding sequences determine the length and stability of the transcript [121–123].

1.3 Methods for genome-wide identification of enhancers

Next-generation sequencing techniques can be used to predict and annotate enhancers in various ways. An analysis may comprise the prediction of enhancers, the determination of cell-type-specific activity and the linking to their target genes [124, 125]. A variable degree of inaccuracy and computational pitfalls inhere in each method, which are outlined in the description for the respective technique. Multiple genome-wide assays can be combined to boost confidence [126], but as the sensitivity of the methods is highly variable, a lot of potential enhancers are erroneously eliminated when the consensus is used only. For example many enhancers require their native chromatin context to function properly [127] and will not show activity in reporter assays as exemplified by the upstream regulatory element (URE) of PU.1 [128–131]. On the other hand even experimental deletion or mutagenesis of a bona-fide enhancer might not result in gene expression changes due to frequent redundancy of cis-regulatory elements [132]. These challenges might explain the large range of estimates on the total number of enhancers in the genome to be found in the literature.

The most commonly used methods for enhancer detection are listed below [reviewed in 133, 134]:

Open chromatin: To function and permit binding of transcription factors or coactivators, all classes of cis-regulatory elements need to be nucleosome-free in their active state. Open chromatin however is also readily accessible for enzymatic digestion by nucelases or integration by transposases. Protocols to map nuclease hypersensitive sites [135] have underpinned the discovery of cis-regulatory elements for decades and were among the first techniques, which were adapted to next-generation sequencing [136]. The assay was widely used during the ENCODE

project [137], but like the related FAIRE-seq (Formaldehyde Assisted Isolation of Regulatory Elements) [138] it is nowadays mostly obsolete. Both have been replaced by ATAC-seq (Assay for Transposase Accessible Chromatin) [139], which features simpler library preparation protocols and can downscale to single cells [140, 141]. All three methods however do not determine the class of the cis-regulatory element in question and require a comprehensive annotation of genes (promoters,exons) for exclusion.

Histone modifications: The nucleosomes flanking an enhancer sequence typically undergo epigenetic alterations, which constitute a relatively distinctive pattern. If ChIP-seq experiments (Chromatin ImmunoPrecipitation) [142–144] for all those modifications are carried out, putative enhancers can be identified in several different states [↪ subsection 1.2.2] at once. The discrimination between active and poised enhancers based on H3K27ac and H3K27me3 respectively [145] is well established, but the guidelines for methylation on lysine 4 are conflicting. A widely cited publication suggested to define enhancers based on presence of H3K4me1 and lack of H3K4me3 [146]. Although both modifications attach to the same lysine residue and are mutually exclusive, each nucleosome consists of two H3 proteins and several nearby nucleosomes can be modified accordingly. Thus, both modifications could technically be detectable close to the same enhancer in parallel. Indeed concomitant H3K4me1, H3K4me2 and H3K4me3 modifications have been reported at enhancers [147, 148] and recent experiments in *Drosophila* embryos and mouse embryonic stem cells (mESCs) indicate that all methylation states are functionally interconvertible [149]. Hence an ambiguous search pattern is a disadvantage of the method, furthermore the size of the nucleosome limits the spatial resolution to about 2 kb (with exceptions [150]), which impedes motif analyses of the core enhancer. On the other hand high quality antibodies to pull down modified histone residues, standardized protocols and kits are commercially available. Thus, such data can be generated with a fair level of expertise from few cells [151] down to single cell level [152].

Coactivator ChIP: If the lineage-determining transcription factors governing the differentiation into a specific cell type are known, their binding to gene-distant sites can guide the characterization of cis-regulatory elements and vice versa [17, 153]. ChIP-seq of cofactors such as the histone acetyltransferase p300 [154] or specific subunits of the mediator complex [155] can also lead to the discovery of new enhancers. Since immunoprecipitation of transcription factors is generally challenging and specific monoclonal antibodies are often unavailable, this approach is rarely chosen to identify enhancers. Furthermore, it has been suggested that transcription factor occupancy does not equal function [156] and false-positive signals are quite common [157].

eRNA profiling: The transcription of eRNAs [↪ subsection 1.2.3] is indicative of active enhancer function. GRO-seq, which assesses nascent transcription [72, 158], as

well as CAGE-seq [159,160], which maps the 5'transcription start sites of mRNAs, can be used to detect enhancers. However, both computational analyses are intricate and potentially prone to a large fraction of false-positives [103,161]. A huge advantage is the promoter-specific simultaneous gene expression profiling [162], which simplifies the identification of enhancer target genes and somewhat compensates for the challenging library preparation. Furthermore the exact location of the core enhancer can be traced with unparalleled precision due to the narrow peaks.

Sequence-based: To a certain extent, sites of enhancers can be derived purely from the underlying sequence, if transcription factor binding motifs are known and reference genomes of some reasonably related species are additionally available for comparison. Although cross-species conservation of individual enhancers is low [163–165], frequent motif patterns and higher-than-average sequence constraints are suggestive of cis-regulatory elements [reviewed in 166]. Computational tools, such as the ENHANCER ELEMENT LOCATOR [167] can be used to screen for genes regulated by certain transcription factors. The opposite analysis is performed by KMER-SVM [168] or HOMER [169], which will detect recurring motifs in a given set of cis-regulatory elements.

Reporter-based: As mentioned previously, some enhancers require their physiological native chromatin environment to function properly [127]. For others this premise is negligible and they can be successfully cloned into reporter constructs for screening. Typically the enhancers sustain the expression of a reporter gene, such as a fluorescent protein or an enzyme to power a photometric or colorimetric reaction [reviewed in 170]. Next-generation sequencing has enabled high-throughput variants of those assays, one of which is STARR-seq [171]. It was gradually adapted to mammalian genomes [172, 173] and features a transcript containing the sequence under investigation. Thus, the enhancer element is contained in the transcript and can be recovered from a RNA sequencing. A drawback of many reporter-based assays is the transcription initiated at the bacterial ORI [174], which is present in many vectors, as well as the triggering of an interferon response in the transfected target cells [175].

The Rosenbauer lab is proficient in the use of four different methods of various categories to detect enhancers: H3K4me1/H3K27ac ChIP-seq, CAGE-seq eRNA detection, STARR-seq and ATAC-seq.

1.4 Enhancer involvement in pathogenesis

The majority of known sequence variants that contribute to disease[1] can be found in ENCODE-annotated cis-regulatory elements. Mutations within those sites may abrogate their function or severely alter their properties, but predicting the consequences in non-

[1] according to the NCBI CLINVAR database, normalized to sequence length

coding DNA is challenging [176, 177]. In contrast, evaluation of a specific nucleotide change in protein-coding DNA is relatively straightforward on the grounds of the known codon table.

Several laboratories have devoted themselves to establish an analogous *cis-regulatory grammar* [178], which is believed to exist, because sequence features in enhancers and the arrangements of transcription factor binding sites often conform to certain physical constraints [179]. The non-coding DNA grammar is clearly neither as universal nor as distinct as the codon code, but various techniques such as THERMODYNAMIC STATE ENSEMBLE MODELS [180] have already contributed to formalize some of the rules. A better understanding of such instructions does not only aid in interpreting disease causing genetic variation in non-coding DNA, but also helps to understand the exaptation of transposable elements into novel cis-regulatory sites [166, 181].

It should be stressed that also aberrant epigenetic modifications may disrupt proper regulation and thus function additionally or jointly with mutations in a pathogenic manner [182].

1.4.1 Examples of pathogenic enhancer abnormalities

Substantial knowledge on the basic principles of genetic regulation was derived from studies at the β-globin gene cluster and so were the first indications for pathogenic cis-regulatory effects. DNA from thalassaemia patients was tested for genetic rearrangements, but sometimes the integrity of the β-globin gene itself was untampered with. In one case study the gene was aberrantly hypermethylated and silenced in vivo due to an unspecified cis-effect [183]. Another case study a few years later demonstrated that a particular 100 kb deletion contained two Alu elements [184], whose enhancer activity could be shown in vitro. Furthermore the abrogation of the locus control region will also lead to thalassaemia [185]. In all named cases the pathogenic phenotype was a result of insufficient β-globin protein levels.

In this sense, the clearest causalities arise from the disturbance of genes that are closely linked to a certain disease. Pre-axial polydactyly, a Mendelian disorder caused by ectopic SHH expression [56] is another example of a straightforward causality. However, in accordance with cell-type or tissue-specific enhancer activity, an enhancer-mediated pathogenesis may not fully recapitulate the clinical phenotype of patients who carry mutations in the genes itself. This is illustrated by the Holt-Oram syndrome: Pathogenic TBX5 mutations will typically present itself by limb malformations and heart defects, whereas a disrupted enhancer may only affect the cardiac expression and permit normal limb development [186].

The contribution of enhancers to multifactorial, non-Mendelian medical conditions is even harder to assess, but might be nevertheless unraveled by association studies. Such a study proved that the susceptibility for Hirschsprungs disease strongly increased by an incapacitated intronic enhancer of the RET tyrosine kinase, despite low penetrance and different genetic effects in males and females [187].

Association studies were formerly conducted mostly between relatives from families with an elevated risk for a certain condition and assayed only few genetic loci. The advent of microarrays and subsequently the era of next-generation sequencing has shifted the attention to genome-wide association studies (GWAS), which screen large patient vs. control cohorts and test thousands of genetic variants in parallel - e.g. for Alzheimer's disease [188].

This unbiased approach allows to detect alterations also in unsuspected distant cis-regulatory regions, but is prone to false-positives due to random associations within the large number of tested regions. Thus, considerable controversy sparked over the benefits and accuracy of GWAS, especially in terms of risk assessment and counseling in personalized genetic testing [189].

Strategies to undergird low-penetrance risk variants include experimental validation by CRISPR-mediated targeted mutation [55] or the quest for subgroup-specificity [190] and recurring association with different diseases. The latter is exemplified by the single-nucleotide polymorphism rs6983267, which lies within a transcriptional enhancer of the MYC proto-oncogene and increases the risk for colorectal as well as prostate cancer [191, 192] due to a higher responsiveness to Wnt-signaling [193]. Nevertheless the determination of enhancer effects on susceptibility to cancer remains challenging [reviewed in 194–196].

1.4.2 Enhancers contribute to leukemogenesis

Hematopoiesis, the development of diverse mature blood cells from hematopoietic stem cells requires an intricate regulation. The appropriate expression of key transcription factors such as PU.1, GATA1, GATA2 or C/EBPα at various stages governs progenitor commitment and differentiation. Ten-thousands of enhancers are presumably involved in hematopoietic regulation in total [151, 197, 198].

Generalizations about leukemogenesis are almost futile, given the many different subtypes . However, in some cases notorious genes like MYC and its enhancers, which are repeatedly implicated in cancerogenesis, do affect other cancers as well as leukemia [199, 200]. In contrast, the downregulation of PU.1 is restricted to hematopoietic cancerogenesis. None the less, it represents a proven route to leukemia [129, 201] and already subtle PU.1 reduction by a heterozygous deletion of an enhancer was sufficient to initiate a myeloid-biased preleukemic state [202].

Especially late-onset leukemia are characterized by the presence of preleukemic hematopoietic stem cells, which have progressively acquired an increasing mutation burden over their lifetime. These cells are not yet leukemic and expansive, but exhibit spurious alterations in their gene expression programs and enhancers. Such genetic regulation patterns mimicking disparate developmental stages, ultimately increase susceptance to uncontrolled cellular expansion and are retained in the descendant leukemic clones, which was revealed by a study in late acute myeloid leukemia (AML) [203].

Unsurprisingly, aberrant super enhancers (respectively LCRs) strongly promote preleukemic states, since they govern the activation of whole gene clusters. This has been elaborated with regard to T-cells and the pathogenesis of T cell acute lymphoblastic leukemia (T-ALL): The introduction of binding motifs for the MYB transcription factor by somatic mutations forms a novel super enhancer upstream of the TAL1 oncogene and sustains its expression [204, 205]. In a particularly dismal ALL subtype driven by TCF3-HLF, the chimeric transcription factor activates an enhancer cluster controlling expression of the MYC gene and instigates the respective transcriptional program [206]. Because hematopoietic MYC expression is intricately regulated by combinatorial and additive activity of individual enhancer modules within this cluster [207], a dysregulation of MYC program can be mediated by various factors or arise from amplifications within the enhancer region [200]. Therefore, the enhancer cluster is complicated in many leukemia subtypes and also pivotal for MLL-AF9-driven AML [207].

A different mode of action has been reported for a distinct subtype of acute myeloid leukemia. In AML with the *inv(3)(q21;q26)* karyotype [208] a genomic rearrangement repositions a distal hematopoietic enhancer of GATA2 in close proximity to the stem-cell regulator EVI1, which is ectopically activated. Concomitantly, GATA2 expression is diminished and both events facilitate leukemic expansion [209, 210].

Supplementary Chapter 2

Methylome data of MLL-AF9 leukemia

2.0 Whole genome bisulfite sequencing

This text precedes section 2.1 and provides additional information about the WGBS sequencing as well as the cross-sample consistency for the replicates

Previous studies have already addressed the effects of Dnmt1 reduction in MLL-AF9 [211,212], but did not generate genome-wide methylome data such as RRBS or WGBS. To obtain the required amount of genomic DNA for WGBS, we had to broaden the c-Kit gate in the published flow-cytometry protocol [213] slightly: Instead of sorting the 10% cells with the highest c-Kit expression (termed c-Kit^{high}) we resorted to a 20% cutoff, which is well in agreement with other protocols [214]. Nevertheless we refer to this population as c-Kit^{+} to avoid confusion. MLL-AF9 Dnmt1$^{+/+}$ c-Kit^{+} cells expressed significantly more Dnmt1 than a c-Kit^{low} (10 % of the cells with the lowest c-Kit expression) control population, suggesting that Dnmt1 is particularly expressed in LSC-enriched fractions. Importantly, we also confirmed the expected lower Dnmt1 expression in both populations from Dnmt1$^{-/chip}$ mice versus the respective wild-type control.

The extracted genomic DNA was sent to our collaborators, the laboratory of Frank Lyko at the German Cancer Research Center (DKFZ) in Heidelberg, where it was further processed, bisulfite-converted and sequenced. The local bioinformatician Günter Raddatz ran the quality checks and aligned the reads to the *NCBI37/mm9* reference genome. We obtained a summary of the data with base-resolution genomic CpG-coordinates, methylation rate (methylscore) and read coverage, which was subsequently used for all further analyses.

As it was considered to be futile to obtain the required amount of cells to generate WGBS data from normal hematopoietic stem cells at that time (advanced protocols requiring much less input material were not yet available), the generation of healthy controls was omitted. For this reason, we accessed a published dataset of the Goodell laboratory, which comprised meta-samples of hematopoietic stem cells (HSCs) from several pooled mice after secondary transplants [215].

A hierarchical single-linkage clustering of the individual samples [▷ Figure S2.1] showed

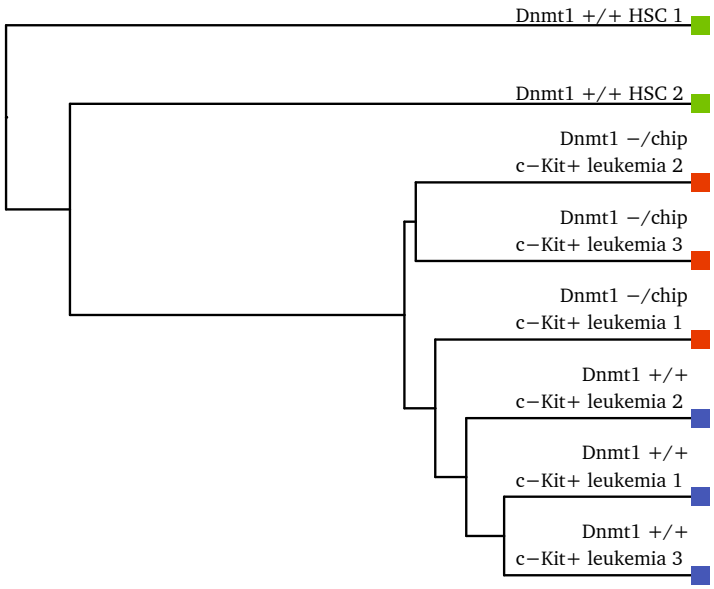


Figure S2.1: Hierarchical single-linkage clustering of the separate WGBS samples. The order was based on the Euclidean distance between the respective methylscores of all CpGs, which were covered with >3 reads in all samples ($n = 65\,184$, 0.15 %).

that our own leukemia specimen clustered away from the healthy hematopoietic stem cells, thus confirming leukemia specific changes to the methylome. Evidently the Kendall rank correlation [▷ Figure S2.2] corroborated leukemia-specific differences, but also shed light on the relatively poor comparability of the two HSC controls. Among the MLL-AF9 replicates, the Dnmt1$^{+/+}$ leukemia showed a high cross-sample consistency, which was to a slightly lesser extent also the case for the Dnmt1$^{-/chip}$ genotype, which may be attributable to variable Dnmt1-levels. However, it should not go unnoticed that these calculations could be skewed due to the small number of CpGs with sufficient coverage in all samples (0.15 % of 43 445 912 CpGs in the *NCBI37/mm9* reference genome).

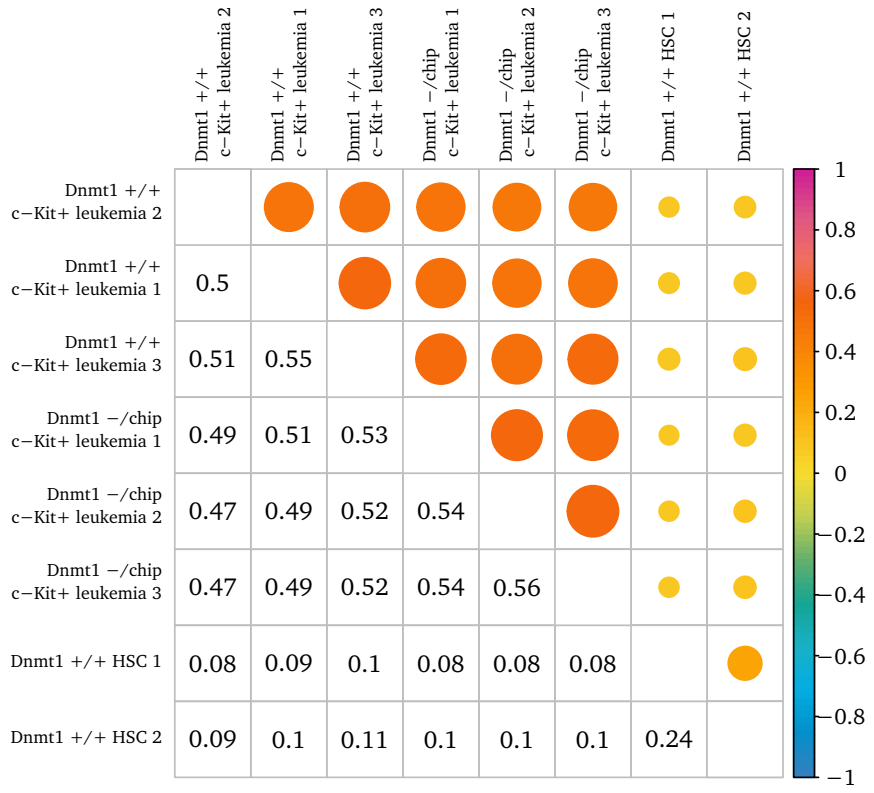


Figure S2.2: Pairwise Kendall rank correlation coefficients τ of the single WGBS samples. Only complete observations, ie CpGs, which were covered with >3 reads in all samples, were considered for the calculation ($n = 65\,184$, 0.15 %).

2.3 Chromatin-state-dependent demethylation

This text supplements section 2.2 and elaborates on the methylation cayons decribed in HSCs. Later studies clearly showed that the canyons are located in open-chromatin intergenic areas and thus are not related to the compromised regions.

2.3.3 Canyon methylation in MLL-AF9 leukemia

In 2014, the group of Margaret Goodell had characterized a possibly new methylation feature in hematopoietic stem cells and termed it canyons [215]. It referred to genomic areas of intermediate size (20 kb - 150 kb), which often encompassed several complete genes or at least their promoters and stood out due to almost absent methylation. Jointly with other groups the notion was established that an intricate balance between de novo methylation and active as well as passive DNA-demethylation shapes the canyons and demarcates their limits [216,217]. Hypermethylation of the canyons can impair leukemogenesis [218] and affects pluripotency and differentiation in general [219].

Therefore, we gave consideration to the possibility that malformed canyons might contribute to the self-renewal impairment observed for Dnmt1$^{-/chip}$ c-Kit^{+} cells. We looked

for eroded borders (due to passive demethylation) or hypermethylation (because of potential increased levels of compensatory de novo methylation) in Dnmt1$^{-/chip}$ without detecting abnormalities [▷ data not shown]. No relevant leukemia-specific aberrations could be found, regardless whether the regular HSC canyons or the enlarged derivatives from Dnmt3a hypomorphic mice were used. On average the methylation of some canyons and the CpG-Islands within slightly increased in Dnmt1$^{+/+}$ leukemia (from 0 to $\leq$0.25), but the vast majority remained fully unmethylated [▷ Figure S2.3]. The intra-leukemia contrast did not point to significant changes in methylation with regard to the canyons. [▷ Figure S2.4]. In summary, we concluded that the deteriorated self-renewal and senescence in Dnmt1$^{-/chip}$ was not related to alterations in the canyons.

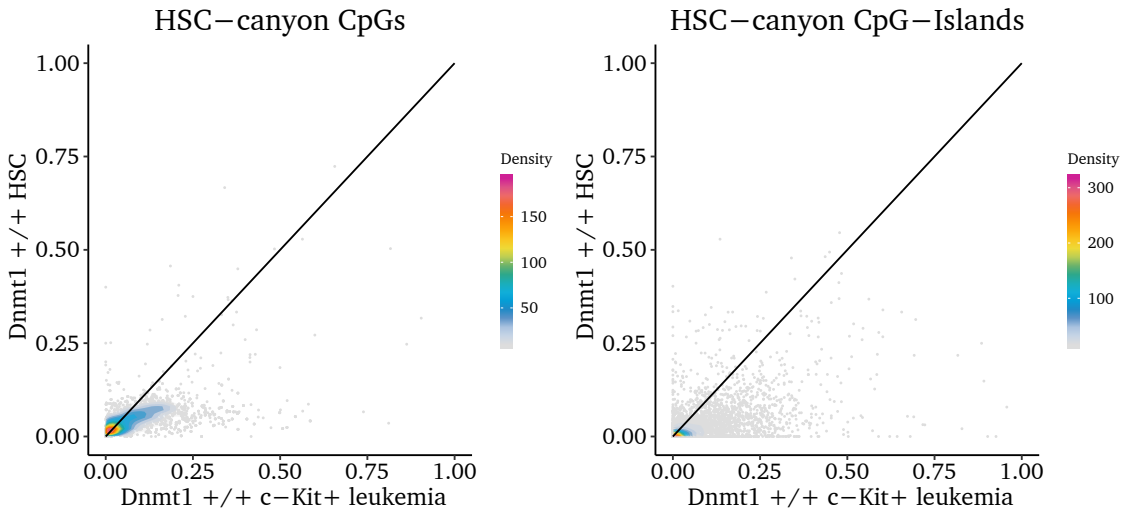


Figure S2.3: Solely CpGs located in known methylation canyons from normal, healthy HSCs were mapped either on 100 kb windows (slid by 25 kb steps) or CpG-Islands to generate these scatterplots of average methylscores.

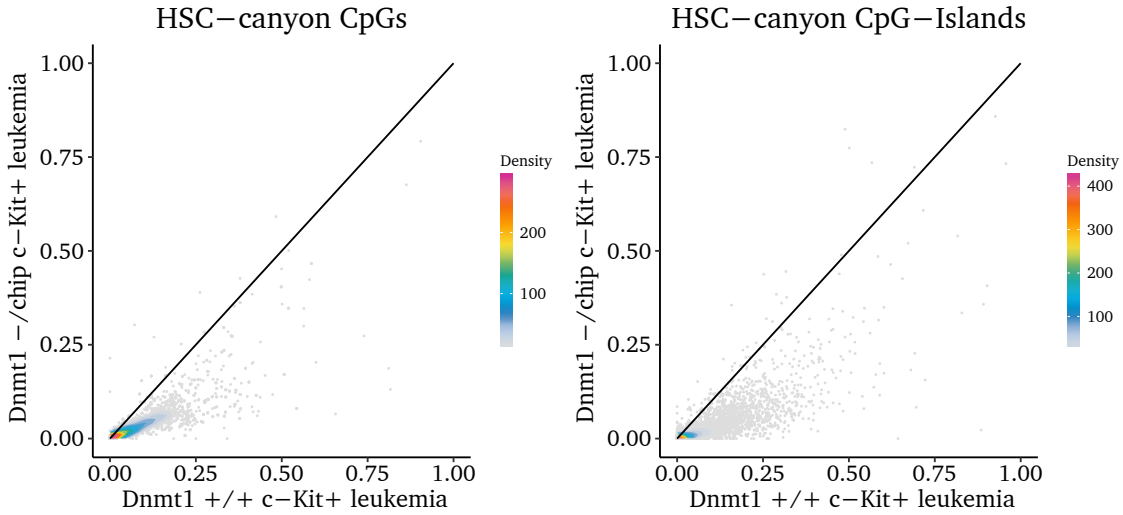


Figure S2.4: Contrast of Dnmt1$^{-/chip}$ vs. Dnmt1$^{+/+}$ c-Kit^{+} leukemic cells for the same regions and CpGs used above for Figure S2.3.

Supplementary Chapter 3

Specification of the compromised regions

This text explains how MethylSeekR works and why the method reached its limits with our methylome data. We also show what we tried to improve to use the software anyway.

3.3 Standard approach failed to discriminate domain borders

Shortly after the first comprehensive WGBS datasets became available, a feature termed *Partially Methylated Domain* (PMD) was described [220]: Some methylomes contained large (>150 kb) regions of seemingly disordered methylation harboring many heterogeneously methylated CpGs.

Since PMDs are associated with AT-rich lamina-associated domains [221, 222], we presumed the applicability of a published tool for PMD calling named MethylSeekR [223] to precisely determine the domain borders of the compromised regions. However, we could not derive meaningful and verifiable limits for the compromised regions with this approach.

One reason may have been the relatively low coverage of less than the recommended 10 x at virtually any genomic region even after pooling the replicates into metasamples. But we also asked, whether the compromised regions in Dnmt1$^{-/chip}$ leukemia are truly equivalent to PMDs in every regard.

Furthermore also the persistent ciLAD regions comprised a decent amount of partial methylation, which may have resulted in a lack of control regions and thus impeded a proper function of the MethylSeekR software.

3.3.1 Emulation of the MethylSeekR methodology

To address this issue, we faithfully recapitulated the approach used by the software, which is outlined in the accompanying paper [223]. Briefly, the MethylSeekR program models the methylscore as being sampled from a probability density function (PDF) of the beta distribution family [▷ Equation 3.1]. Usage of the beta distribution family to delineate methylscores is a common proposition in modeling methylome data [224–226]. The shape of the curve is determined by two parameters designated $\alpha > 0$ and $\beta > 0$, but

the desired upward open parabola will only result when $0 < \alpha < 1$ and $0 < \beta < 1$, which is therefore the most applicable range for methylome data [▷ Figure S3.1, left panel]. METHYLSEEKR restricts itself to symmetric beta functions and equalizes $\alpha = \beta$. Hence the probability f_i of a given CpG i to be methylated is modeled by METHYLSEEKR as shown in Equation 3.2 [223].

$$f(x) = \frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{3.1}$$

$$\mathrm{P}(f_i|\alpha) = \frac{1}{\mathrm{B}(\alpha, \alpha)} f_i^{\alpha-1}(1-f_i)^{\alpha-1} \tag{3.2}$$

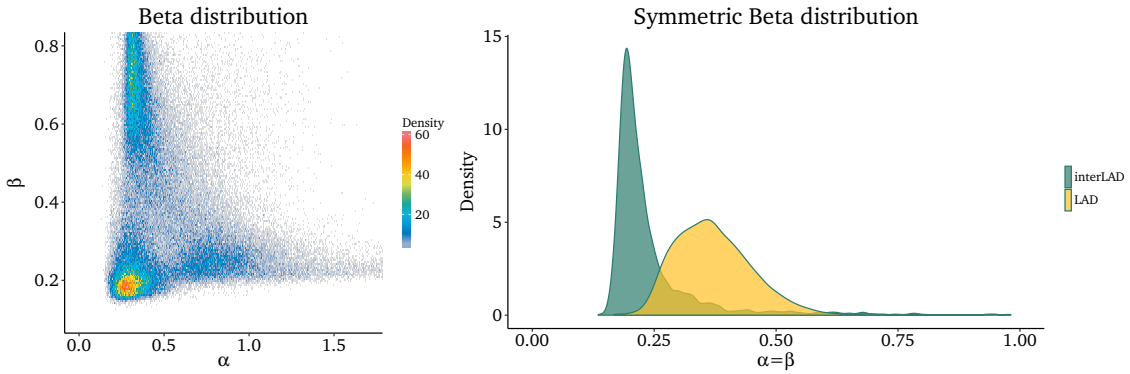


Figure S3.1: Density plots of the parameters α and β for beta regression fits on windows of 100 adjacent CpGs for IMR90 fibroblast methylomes. PMDs appear mostly as regions with $\alpha < 0.5$ and $\beta > 0.6$. In the right panel the genomic regions were annotated with measured lamina-association data [227]. Lamina-associated (LAD) and open inter-lamina regions(iLAD) were treated separately for plotting, highlighting the bimodality of the distribution, which forms the basis for the mixture model of METHYLSEEKR. Mind the long-tailed part of the iLAD density, which is indistinguishable from the LAD values for any modeling approach and will be assigned falsely.

In the subsequent step, METHYLSEEKR fits a mixture model on the resulting density over α, which typically assumes a bimodal distribution split between open and lamina-associated chromatin. Usually PMD-like areas exhibit values of $\alpha > 1$ for symmetric fits. The latter is illustrated, when the methodology is applied on published methylome data from IMR90 fibroblasts, which is known to harbor prominent partially methylated domains [▷ Figure S3.1, right panel]. The Mixture Model treats a particular α-value as being sampled from one of two distributions, which are estimated from the data, but are considered as known ground truth. The algorithm decides, which distribution explains the measured α-value better and joins adjacent regions with the same assignment into blocks to call PMDs.

3.3.2 Improvements to the MethylSeekR methodology

A caveat with the METHYLSEEKR approach is that methylscores frequently assume the extremes 0 and 1, which cannot be simulated in the published way by beta regression

models. The density of beta distributions with the parameters $0 < \alpha < 1$ and $0 < \beta < 1$ is infinite at zero and one, so they are generally only defined in the open interval $]0,1[$. To address this issue, we resorted to a prior transformation of f_i.

$$t_i = \frac{f_i \cdot (n-1) + 0.5}{n} \tag{3.3}$$

The methylscores f_i of n adjacent CpGs within a genomic region were transformed according to Equation 3.3 [228]. The transformed variable t_i assumed extreme values slightly above zero and below one respectively and thus a beta regression model could be fitted on the values.

As a second improvement we also fitted the Kumaraswamy distribution, a beta-type distribution with more convenient tractability [229,230]. As the Kumaraswamy distribution, whose probability density function PDF is given in Equation 3.4 is defined within the interval $[0,1]$, a prior transformation was not necessary. In contrast to the cumulative distribution function (CDF) for a beta distribution, which cannot be reduced to elementary functions unless its parameters α and β are integers, the CDF of the Kumaraswamy distribution has a simple form [▷ Equation 3.5]. This simplicity makes it possible to derive tangible conclusions about the methylscores within a genomic region for known parameters a and b. Furthermore, the CDF can easily be compared to the empirical distribution function (ECDF) of the interval with a one-sided KolmogorovSmirnov test. Because of these advantages, we advocate the Kumaraswamy distribution to replace the widely used beta distribution [223–226] in modeling methylome data.

$$f(x) = a \cdot b \cdot x^{a-1} \cdot (1 - x^a)^{b-1} \tag{3.4}$$

$$F'(x) = 1 - (1 - x^a)^b \tag{3.5}$$

3.3.3 Maloperation of both beta-like models

Nevertheless, neither the original MethylSeekR approach nor the Kumaraswamy fits resulted in a reasonable classification and both failed to correctly resolve the borders of the compromised and persistent regions in Dnmt1$^{-/chip}$ methylation data. Already the first step, the fitting of beta-like functions, turned out to be deficient: Neither for windows of 100 adjacent CpGs (like MethylSeekR uses them) nor for larger sections the beta-like functions could be reasonably fitted [▷ Figure S3.2].

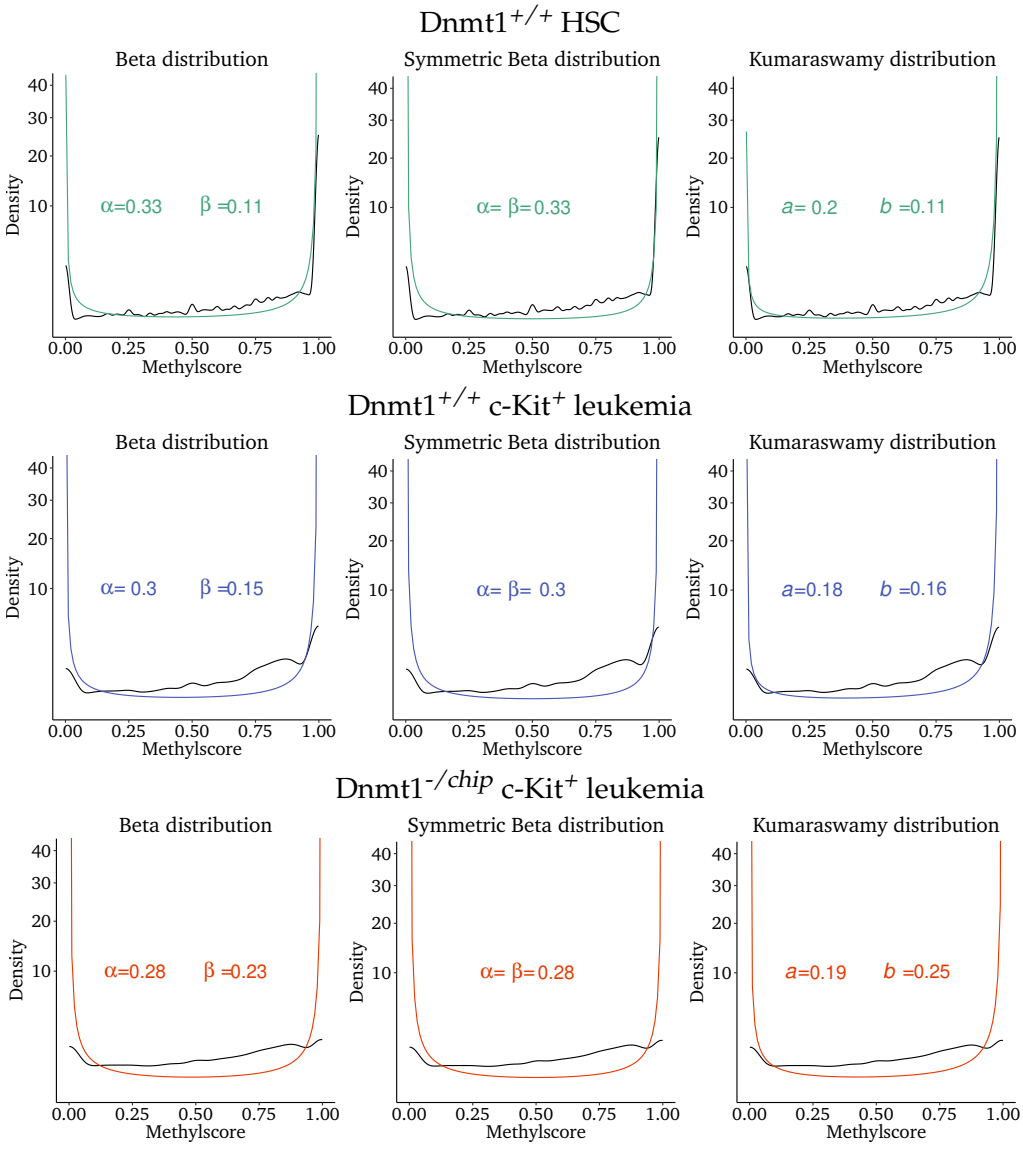


Figure S3.2: Visualization of the best fits for the respective functions (columns) and methylomes (rows). The black lines indicate the real density and the colored lines retrace the fitted functions with the optimized parameters.

In particular the fitted beta-like functions failed to accurately represent the accumulation of methylscores within the range of 0.6 to 0.9, which was most evident in Dnmt1$^{+/+}$ c-Kit^{+} leukemia, but also occurred in Dnmt1$^{-/chip}$ [▷ Figure S3.2, middle and bottom row].

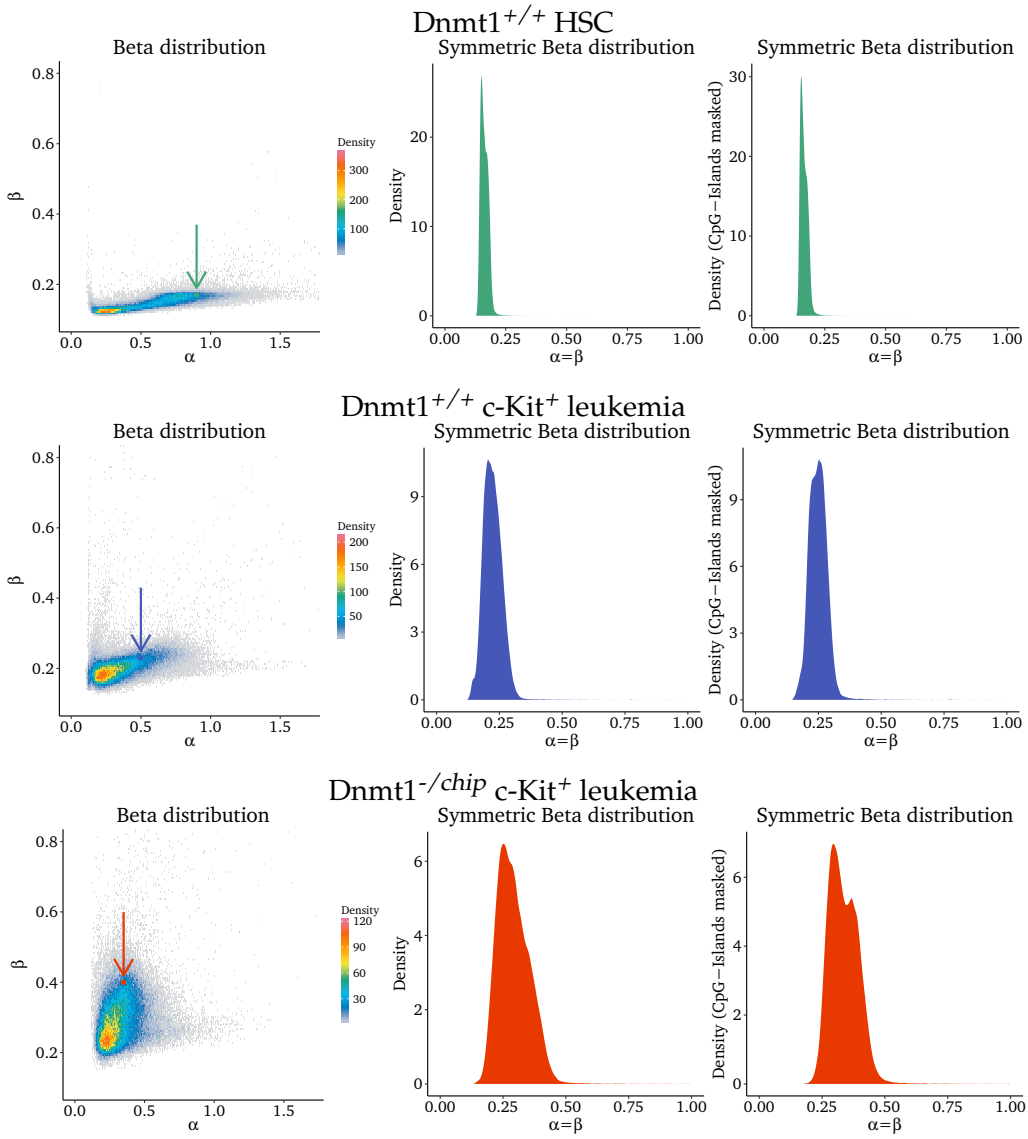


Figure S3.3: Density plots of the parameters α and β for beta regression fits on windows of 100 adjacent CpGs. The left column depicts the 2D density for a regular two-parameter function, whereas the center and right column show symmetric fits only. The figures in the rightmost column show the densities for α-values calculated after prior exclusion of CpG-Islands. The colored arrows highlight α/β parameter combinations, which are rather distinct for a particular sample. The accompanying graphs are shown in Figure S3.4.

While regular beta family fits, which allowed for different α and β values still produced somewhat distinct 2D densities, METHYLSEEKR's self-imposed restriction to symmetric beta functions turned out to be a burden for correct PMD classification. The latter was particularly evident in comparison to the published IMR90 fibroblast methylome [▷ Figure S3.3 vs. Figure S3.1]. In the fibroblasts, PMDs appeared mostly as regions with $\alpha < 0.5$ and $\beta > 0.6$, which were rare (in leukemia) or absent (in HSCs). While the general trend towards smaller α values and larger β values from Dnmt1$^{+/+}$ HSC over Dnmt1$^{+/+}$ c-Kit^{+} to Dnmt1$^{-/chip}$ c-Kit^{+} was still intact [▷ Figure S3.3, left column], the beta distribution was insufficient to capture the slight increase in partial methylation [▷ Figure S3.4], as it was less prominent than in fibroblasts.

This at least partial distinctness was fully lost, when only symmetric functions were used. For leukemia, the α density did exhibit a slightly increased skewness to the right, but remained essentially unimodal, even when the more persistent CpG-Islands had been excluded [▷ Figure S3.3, center and right column]. Thus, the Mixture Model could not be fitted accurately and METHYLSEEKR failed on our methylome data to classify PMDs.

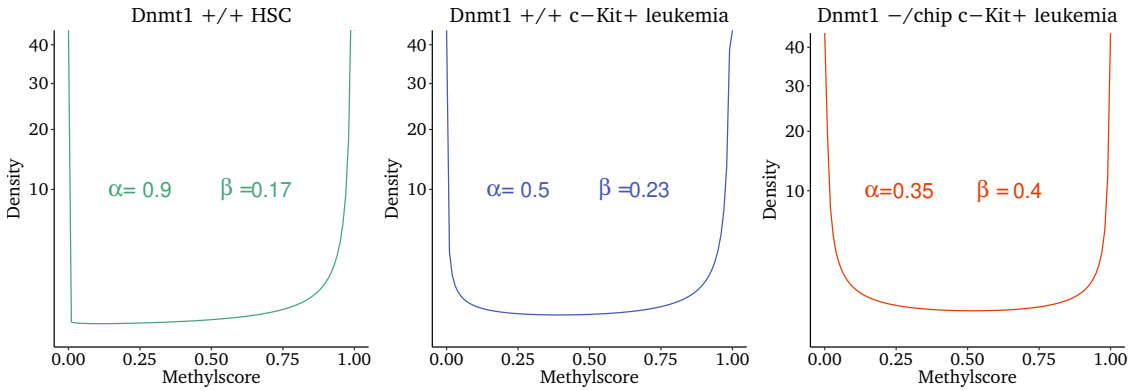


Figure S3.4: Graphs of rather distinct beta PDF parameter combinations, which are significantly more common in the respective sample than in others. The arrows in the left column of Figure S3.3 point to the particular parameter value pair.

Therefore, we developed and evaluated other classification approaches for our data. The one which we ultimately used was based on a Generalized Additive Model (GAMs), but we had also unsuccessfully tried to model the persistency as time-discrete Markov process with fixed transition probabilities between persistent and compromised states [▷ data not shown].

Supplementary Chapter 4

Modeling the methylation probability

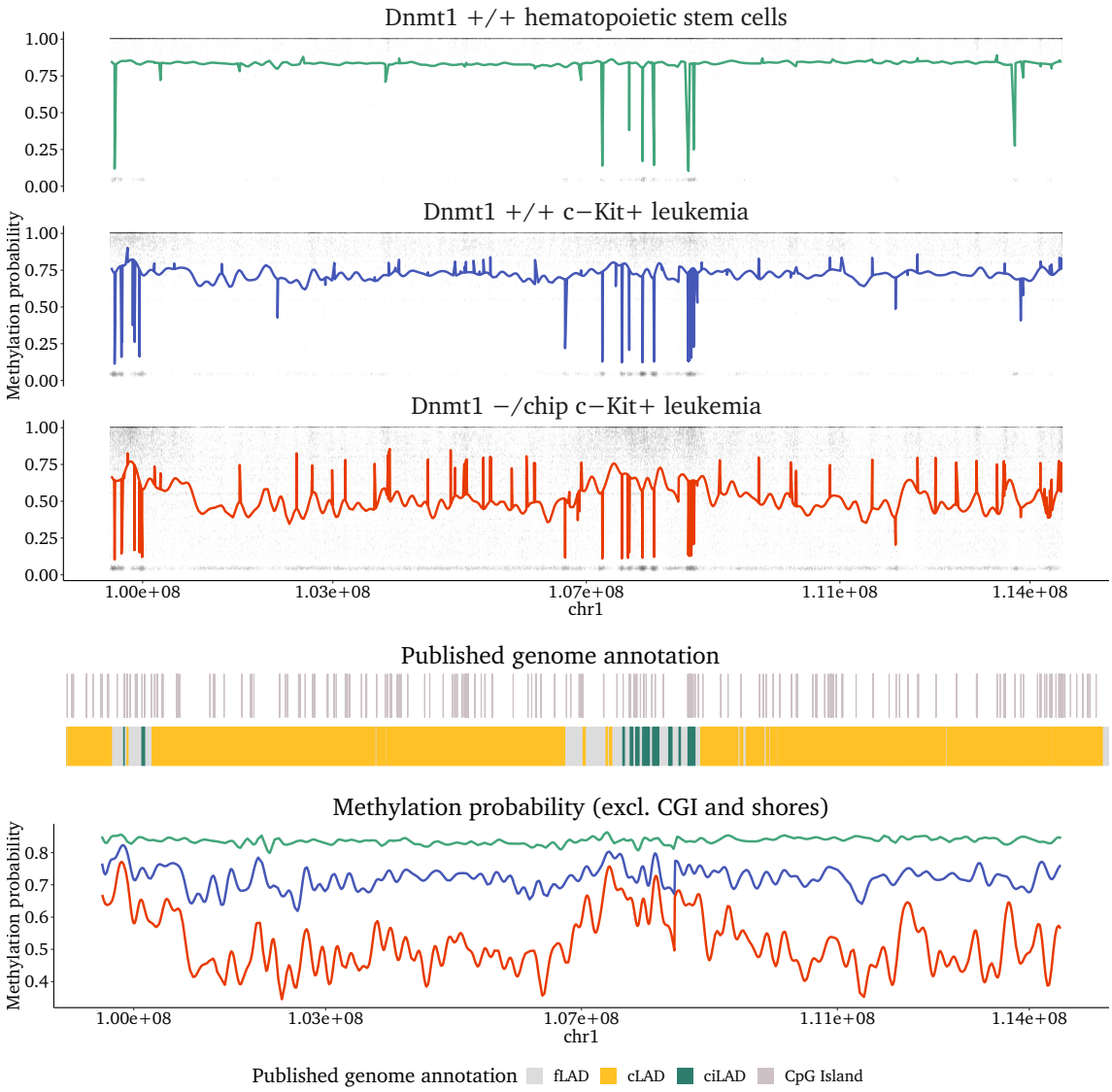


Figure S4.1: Another 1×10^7 bp chromosomal example region. Shown with colored lines is the modeled methylation probability, which is displayed on top of the measured methylation rate of single CpGs in a region. Tiles represent the underlying single CpG data to avoid overplotting - dark hues indicate a high CpG-density. Open chromatin ciLAD regions clearly associate with a higher backbone methylation in Dnmt1$^{-/chip}$ c-Kit^{+} and also harbor most of the demethylated CpG-Islands.

4.3 Comparison of GAM versus a sliding window approach

Smoothness in areas of low coverage and the possibility to account for specific CpG-Island methylation were the two main reasons to fit a GAM. To illustrate this, we conducted a series of comparisons with 500 bp sliding windows.

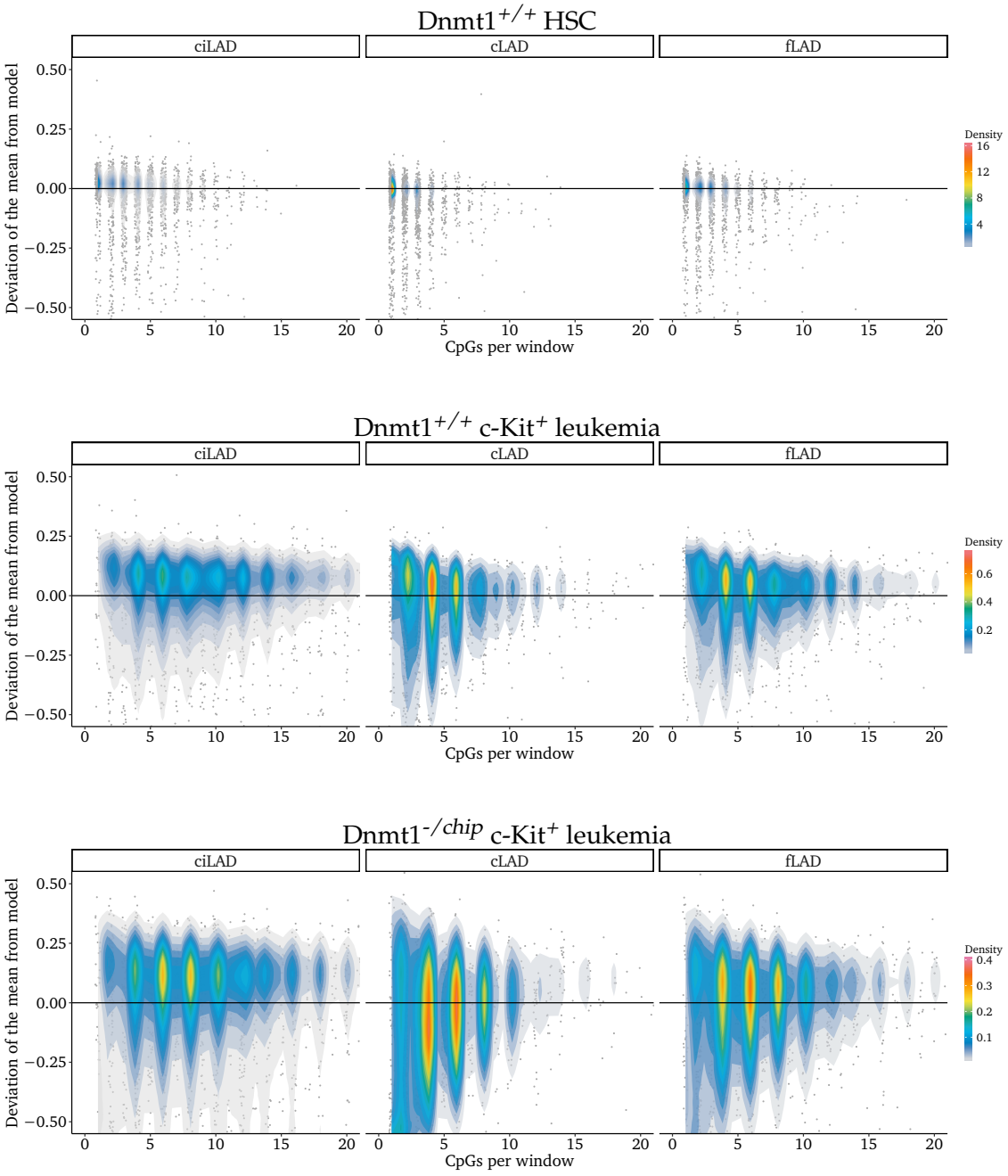


Figure S4.2: Comparison of the GAM-predicted methylscore at the center of a 500 bp window with the average methylscore of all contained and covered CpGs. If the average methylscore surpasses the model's prediction (set to 0), the y-value of the window is positive, while a negative y indicates areas with less than expected methylation. A color-encoded density scale highlights areas with many individual points.

Initially, we compared the average methylscore of measured CpG values within the window to the modeled prediction. For a homogeneous methylome devoid of partial methylation like that of Dnmt1$^{+/+}$ HSCs, the model was generally representative in case the sections comprised 5 or more CpGs and also performed well for the majority of windows with fewer CpGs [▷ Figure S4.2, top row].

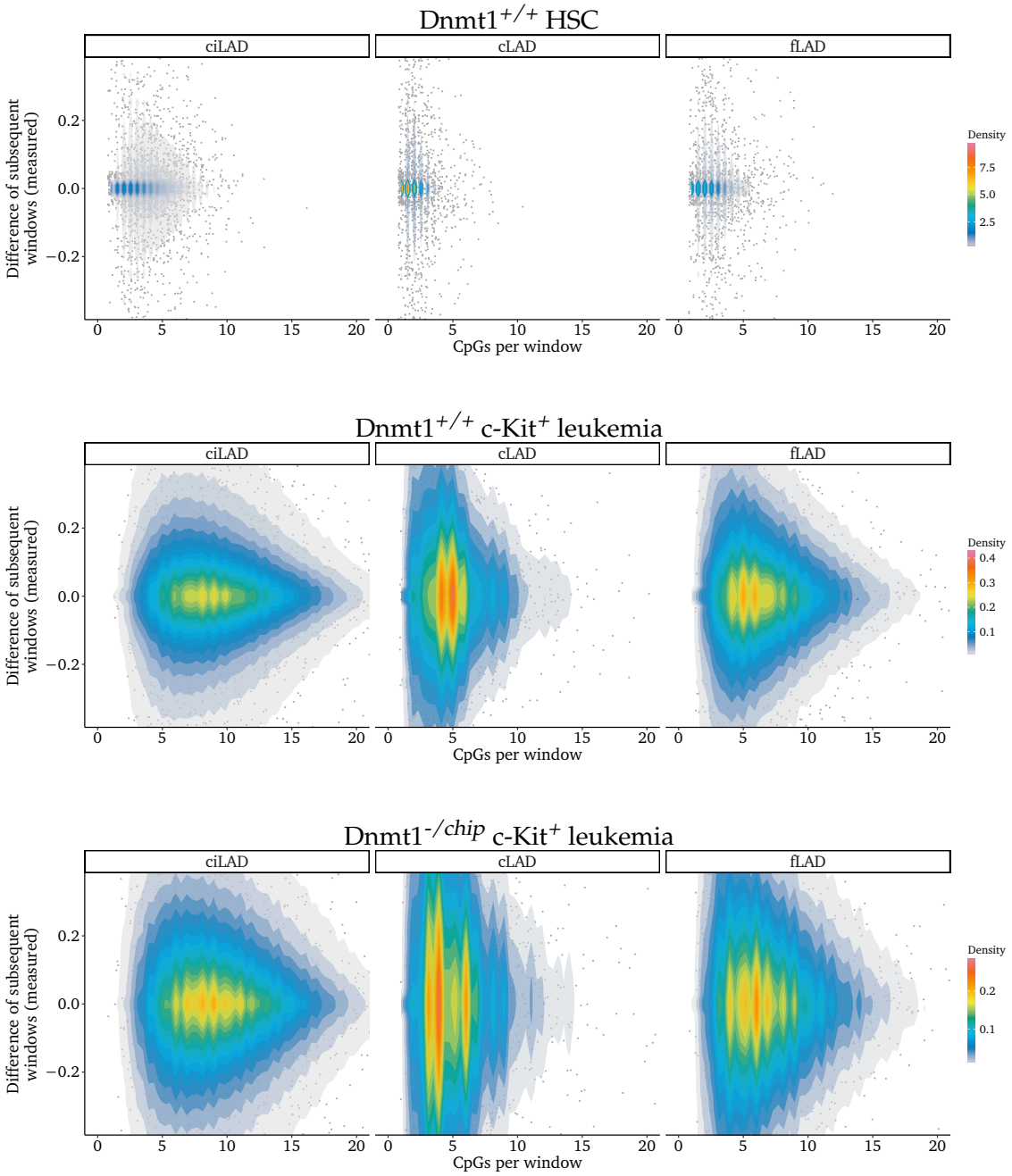


Figure S4.3: Difference of the measured methylscore averages of two adjacent 500 bp windows plotted against the average number of CpGs in these windows. If the downstream (3′) located window is hypomethylated, the y-value is positive, while a negative y denotes pairs with a hypermethylated downstream window.

The deviation of measured and modeled methylation was much greater for leukemia methylomes and decreased only for windows containing more than 10 CpGs. In particular cLADs, which comprised many partially methylated CpGs, were difficult to model accurately and their persistency was generally overrated. Contrastingly the methylscore in ciLADs was typically predicted slightly lower than measured, which also held true for fLADs to a lesser extent [▷ Figure S4.2, middle and bottom row]. Thus, the model was a reasonable compromise that could do justice to all three categories.

Nevertheless its predictions should be used with caution for small genomic areas due to the model's smoothness, which by far surpassed those of the measured methylation. The latter was illustrated by the disparate average methylation of two subsequent 500 bp windows. Usually two adjacent windows merely differed by less than 0.1 in HSCs, but often by 0.2 or more in leukemia [▷ Figure S4.3]. In part these differences were seemingly attributable to low coverage, because they decreased in ciLADs and fLADs for windows comprising >10 CpGs. Yet in cLADs methylscore averages of neighboring windows were highly variable regardless of the coverage, arguing for a mostly unordered, partial methylation in leukemia [▷ Figure S4.3, center column, middle and bottom row].

Supplementary Chapter 5

Relationship of chromatin structure and methylation persistency

5.1 Categorical methylation persistency

To understand the mechanisms, which lead to the impairment of self-renewal in Dnmt1$^{-/chip}$ leukemia and potentially to senescence, we aimed at the accurate determination of methylation persistency, which was achieved with the GAM model. This model permitted to derive a quantifiable methylation persistency at any given chromosomal position, but on top of that we were also interested in a categorization into regions of rather persistent and compromised methylation to e.g. integrate with gene expression data.

To separate the genome into these two categories, we predicted the methylation likelihood from the fitted GAMs at each reference point of a 500 bp grid superimposed over the mouse genome. Grid points within CpG-Islands or an additional 2 kb transition zone margin (in the literature referred to as shore) were neglected. If the difference between Dnmt1$^{+/+}$ and Dnmt1$^{-/chip}$ exceeded an absolute of 0.158, the specific point was classified as being compromised. This cutoff corresponded to the median demethylation in fLADs [231], which we believed to comprise both compromised and persistent regions. Adjacent grid reference points assigned to the same category were united to larger domains, whereas the midpoint of deviant assignments marked a domain border. To account for borderline cases and reduce cluttering, a state switch was only triggered if the transgression lasted for at least 20 subsequent grid points (hence 10 kb). Smaller state switches were permitted, in case two or more grid points significantly transgressed (> 0.258 for compromised regions, < 0.058 for persistent regions). We retracted resulting assignments in areas of low coverage. Using the BEDTOOLS [232] suite, we employed *merge -d 50000* and combined CpG sites covered in our source data, if they were less than 51×10^4 base pairs apart. Subsequently, we created the *complement* of the merger and used *substract* to revoke predictions in those areas, which where usually located at the distal ends of the chromosome or at long stretches of repetitive DNA. As result, we obtained an annotation of methylation inferred regions, which comprised almost 97 % of the genome [▷ Table S5.1].

Type	$\sum$ bp	$\sum$ %
Persistent region	1.421×10^9	53.69
Compromised region	1.141×10^9	43.16

Table S5.1: Properties of methylation-derived chromatin states. The remaining 3 % were retracted due to insufficient coverage.

An individual domain was typically between 100 kb to 300 kb in size [▷ Table S5.2] and thus highly comparable to the published reference sizes of partially methylated domains (PMDs) [220–222] as well as lamina-associated domains (LADs) [227] . However, for each of the respective domains about a quarter was notably smaller (just 10 kb) and often comprised single transcripts or regulatory regions. It was obvious that such small compromised regions bore resemblance to lowly methylated regions (LMRs) [233], although we did not formally test their true equivalence.

Type	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
PR	724	97.81×10^3	1.761×10^5	3.931×10^5	4.811×10^5	6.11×10^6
CR	31×10^3	78.51×10^3	1.351×10^5	3.151×10^5	3.041×10^5	9.371×10^6

Table S5.2: Summary of persistent region (PR) and compromised region (CR) sizes, which we derived from the modeled methylation probability.

At large, we could confirm both the association of persistent regions with open inter-lamina chromatin as well as the intensified loss of methylation in lamina-associated heterochromatic sections. However, approximately 20 % of the genome behaved in an unexpected way and broke the general rule. We detected two classes of these inconsistencies:

- Firstly, we identified small , divergent regions inside contrary larger homogeneous annotations. Figure S 5.1 features an example of three unexpected persistent regions (highlighted by a black arrow) inside a large compromised region at position 1.03×10^8 bp.
- Secondly, we recorded cases, where the transition between a compromised and a persistent region was shifted relative to those of the annotated lamina-association. It appeared as if the domain border was moved in leukemia, which was comprehensible due to the annotation not fully representative of the leukemia situation. Figure 4.2 in main text depicts such an unexpected region roughly at 10×10^7 bp, where a large persistent region extends far into a lamina-associated cLAD region (marked by a black arrow).

5.1.1 Methylation in unexpected regions

Next, we addressed the possibility that the deviances in methylation persistency could have been incorrect predictions of the model. Therefore, we derived the methylation probability for every CpG site covered by the WGBS data from the GAM models, to compare the modeled and measured values. We separated the CpGs according to the

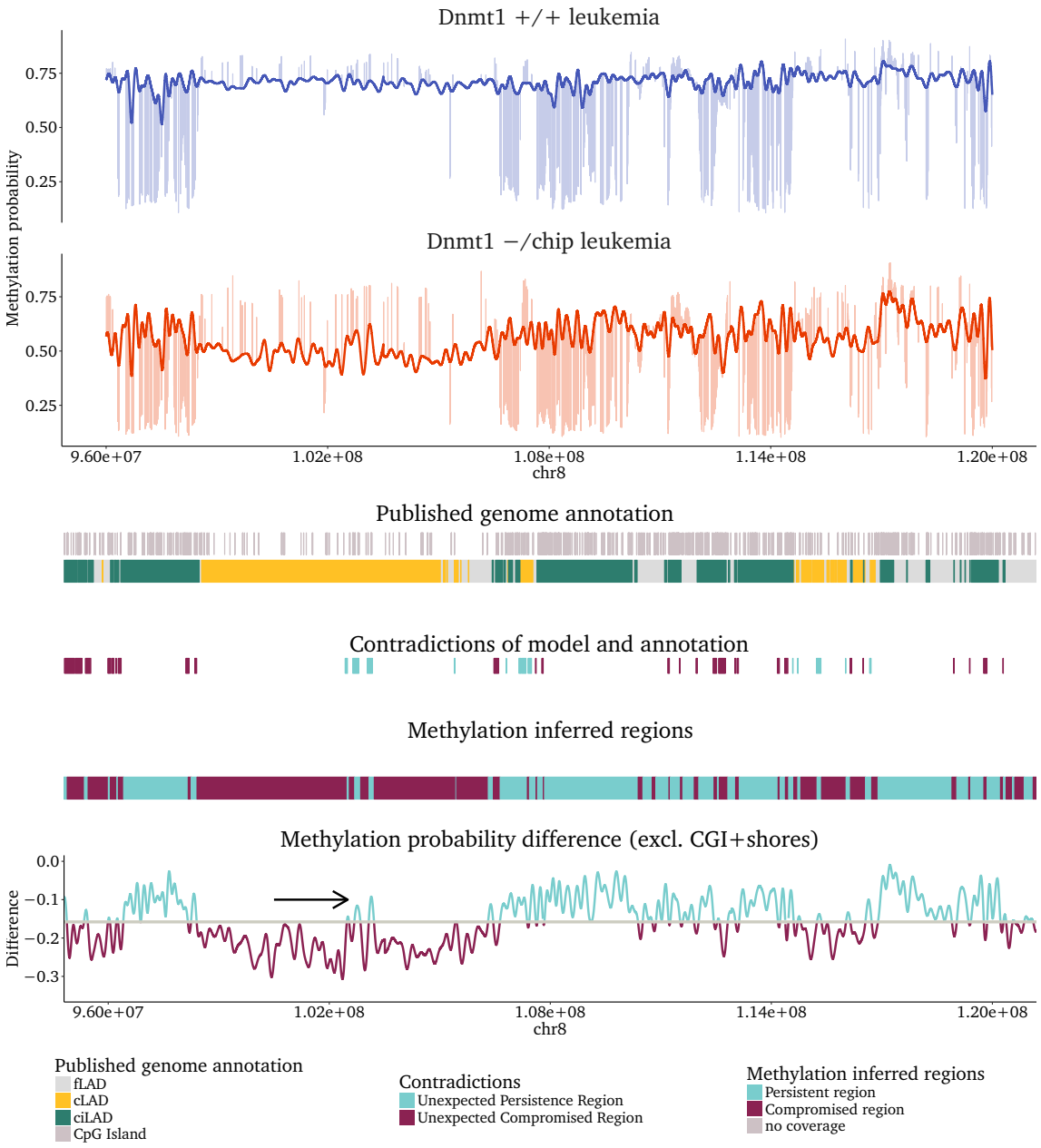


Figure S5.1: Modeled methylation probability of the backbone (saturated lines) and CpG-Islands (pastel lines) in a 2.5×10^7 bp region on chromosome 8. Colored blocks below indicate the extent of chromatin or sequence features on the underlying DNA (ciLADs, cLADs and CGIs). The lower part of the figure shows the categorical methylation persistency with blocks for persistent and compromised regions colored in turquoise and tyrian purple respectively.The lowermost curve retraces the difference between Dnmt1$^{-/chip}$ and Dnmt1$^{+/+}$ c-Kit^{+} leukemic methylation probability, which was used to infer the categorical regions. A black arrow indicates the three unexpected persistent regions mentioned in the main text.

categorical classification into either expected or unexpected and either compromised or persistent regions, respectively. Then, we plotted the densities of the modeled [▷ Figure S5.2] and measured [▷ Figure S5.3] difference.

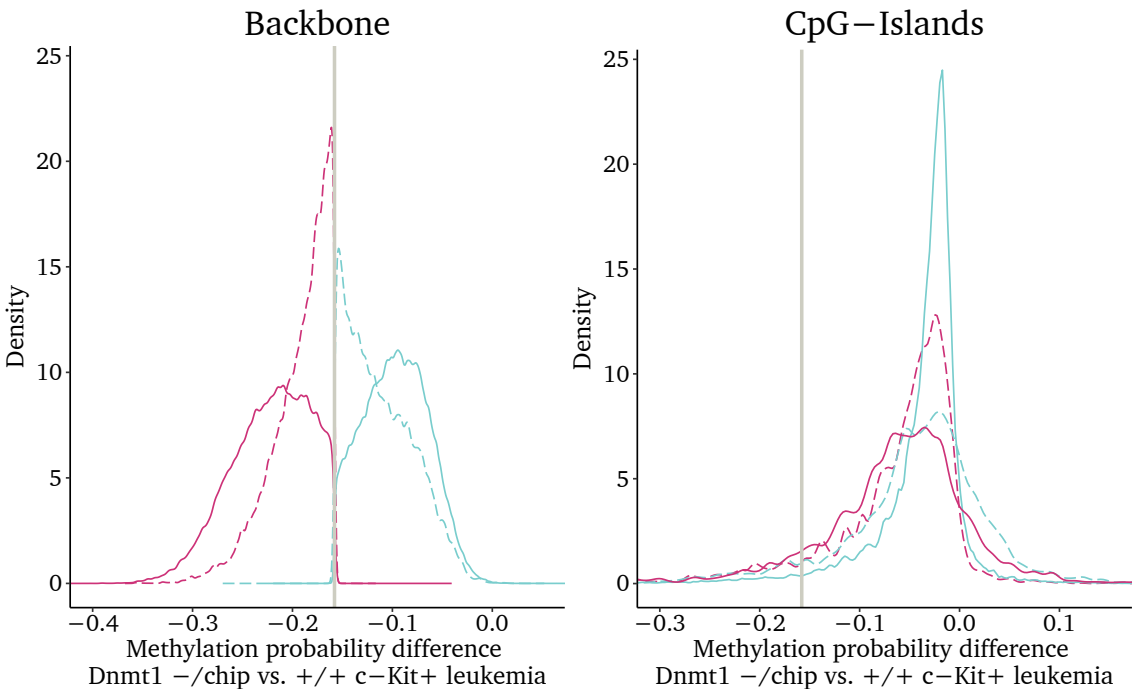


Figure S5.2: Kernel density estimates for methylscore differences on modeled on single CpG level, separated by the GAM-derived persistency categories. A gray vertical line symbolizes the cutoff of -0.158 used to discriminate persistent (represented in turquoise) and compromised areas (colored in tyrian purple). The line type indicates the expectancy, since a solid line marks regions in accordance with the previously published lamina-association [231], whereas the dashed line is chosen for unexpected regions. In the left panel CpGs outside CpG-Islands and their immediate proximity (2 kb, referred to as shore) are shown, in the right panel only CpGs inside the CpG-Islands. A high density implies that many CpGs exhibit a particular modeled probability difference, which is calculated by subtraction of the Dnmt1$^{-/chip}$ c-Kit^{+} methylscore from those of Dnmt1$^{+/+}$ c-Kit^{+} leukemia.

Because of the considerable smoothing, the GAM model did not account for varying methylation probability of individual CpGs. This low volatility meant that virtually all predictions for individual backbone CpGs located in compromised regions resulted in values below the chosen cutoff of -0.158. Vice versa, persistent regions were predicted to exclusively harbor CpGs with a methylation probability difference above the cutoff. Both unexpected regions (dashed lines) clearly represented an intermediate population, which were less compromised or persistent than the respective expected regions [▷ Figure S5.2, left panel]. The density plots for CpG-Islands reflected our previous result that they were globally much more persistent than the backbone. Intriguingly the unexpected compromised regions still comprised the second most persistent CGIs [▷ Figure S5.2, right panel].

The measured methylation data confirmed the modeled ranking for the CpG-Islands perfectly, although CGI persistency in the unexpected compromised regions markedly surpassed the models predictions [▷ Figure S5.3, right panel]. It should further be noted that CGIs even in the least persistent region (the expected compromised regions) harbored on average twice as many stably methylated CpGs than the best maintained sections of the

backbone [▷ Figure S5.3, left panel]. This finding was even more remarkable, considering that CpG-Islands in the compromised regions were typically fully methylated and illustrated the extraordinarily high persistence of CGIs. By and large, the four regions differed only in the number of CpGs that remained unchanged, an observation that we had already made for the lamina association plot.

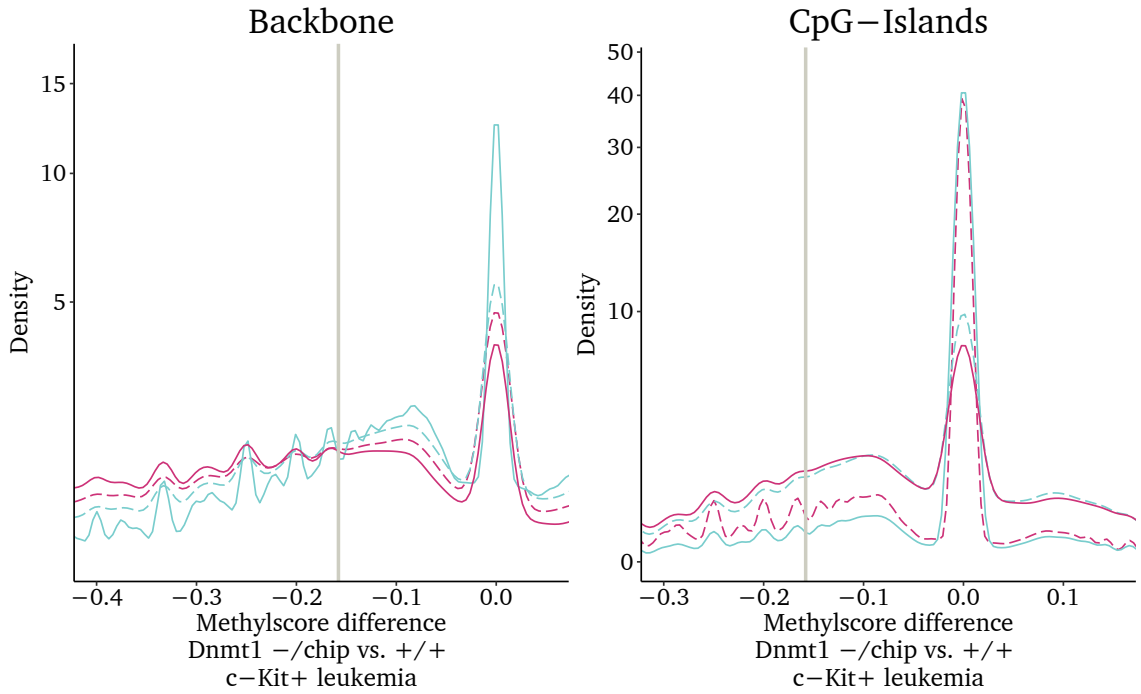


Figure S5.3: Overlay of the density estimates for measured methylscore difference for single CpGs, separated by the GAM-derived persistency categories. Compromised regions are shown in tyrian purple, persistent areas in turquoise. The dashed lines retrace the densities in the unexpected regions in contrast to the solid lines that constitute the expected sections. Mind the different scale (modulus transformed vs. linear), which was needed to display the non-smoothened measurement data, when comparing to Figure S5.2 . The vertical line indicates the cutoff of −0.158 to distinguish persistent and compromised areas.

5.2 Determining factors of methylation persistency

Given that the Dnmt1$^{-/chip}$ genotype exhibits an impaired methylation maintenance, we assumed a mostly passive demethylation. However, we wanted to break down, whether a preferential recruitment of Dnmt1 to specific genomic sites, an unequal de novo methylation by other methyltransferases or selective pressure on functionally relevant CpGs was the driving force. Therefore, we searched for further associations.

5.2.1 Underlying DNA sequence

Already one of the first comprehensive whole-genome bisulfite sequencing studies had suggested that only a fraction of CpG methylation changes occur as part of coordinated regulatory programs [234]. This opened up the possibility that not all methylation is constantly subject to tight control and some may be shaped by chance. Along this line Gai-

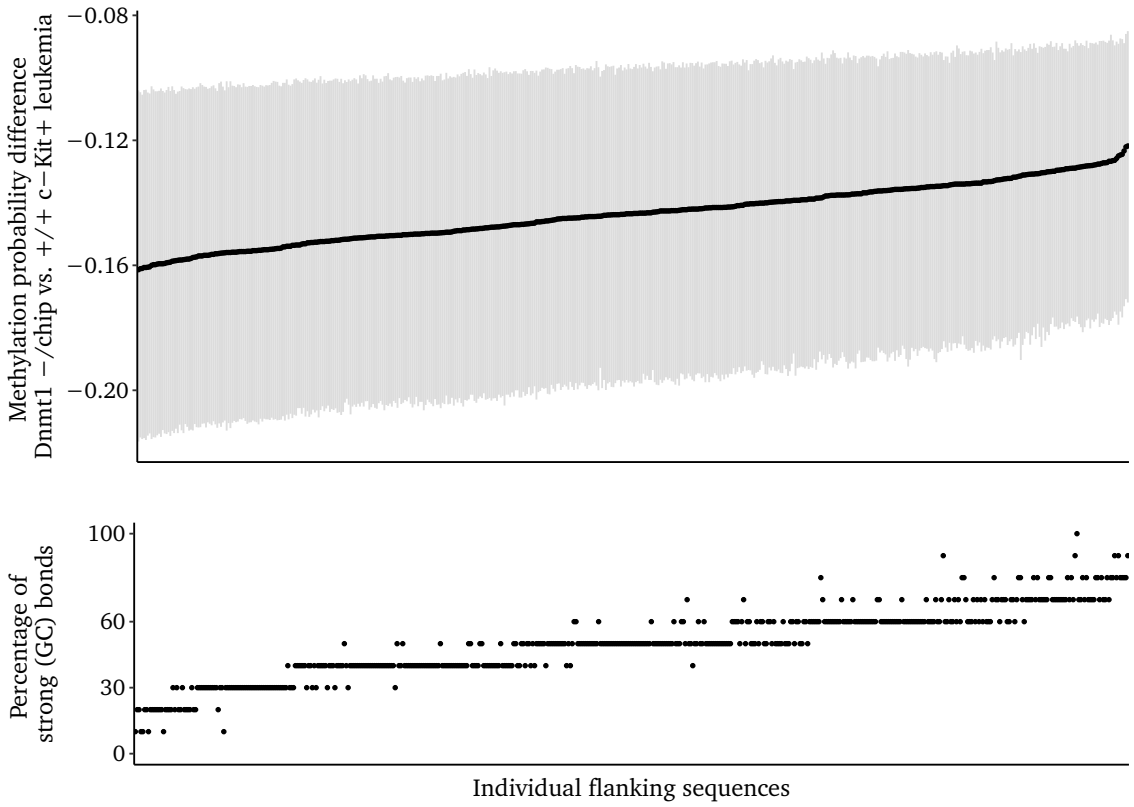


Figure S5.4: The GAM-derived methylation probability for 500 bp grid points is shown in relation to the DNA sequence properties of the surrounding 10 bp window. Shown are the 528 distinct flanking sequences on the X-axis ordered from left to right by increasing methylation persistency. The top panel depicts the median persistency in black and the interquartile range (IQR) in gray , while a dot for each sequence in the lower panel indicates the respective percentage of strong bonds. Typically a higher persistency is linked to CG-content.

datzis and colleagues had proposed that DNA sequence is a major determinant of methylation states outside regulatory regions and in partially methylated domains (PMDs) in particular [222]. Therefore, and due to the similarity of our compromised regions [↪ section 5.1, p.31] with PMDs, we probed their findings on our data.

We asked, if the methylation probability would relate to specific sequence features in the vicinity and tested flanking windows of various sizes. A window of 5 bp in each direction already comprised $4^{10} = 1\,048\,576$ possible sequences, many of which occurred never or only a few times in the dataset. To derive meaningful comparisons, we therefore summarized similar windows. For example, we joined sequences with their respective reverse complement as their methylation persistency was highly similar. Furthermore we subsumed the bases as *S* (strong, G or C) and *W* (weak, A or T), based on our finding that sequence similarity assessed by a variety of string distance metrics (Hamming, Levenshtein or restricted Damerau-Levenshtein) did neither predict the modeled nor the measured demethylation more accurately than the number of strong bonds alone [▷ data not shown].

By and large, we could corroborate the findings of Gaidatzis and colleagues [222] and

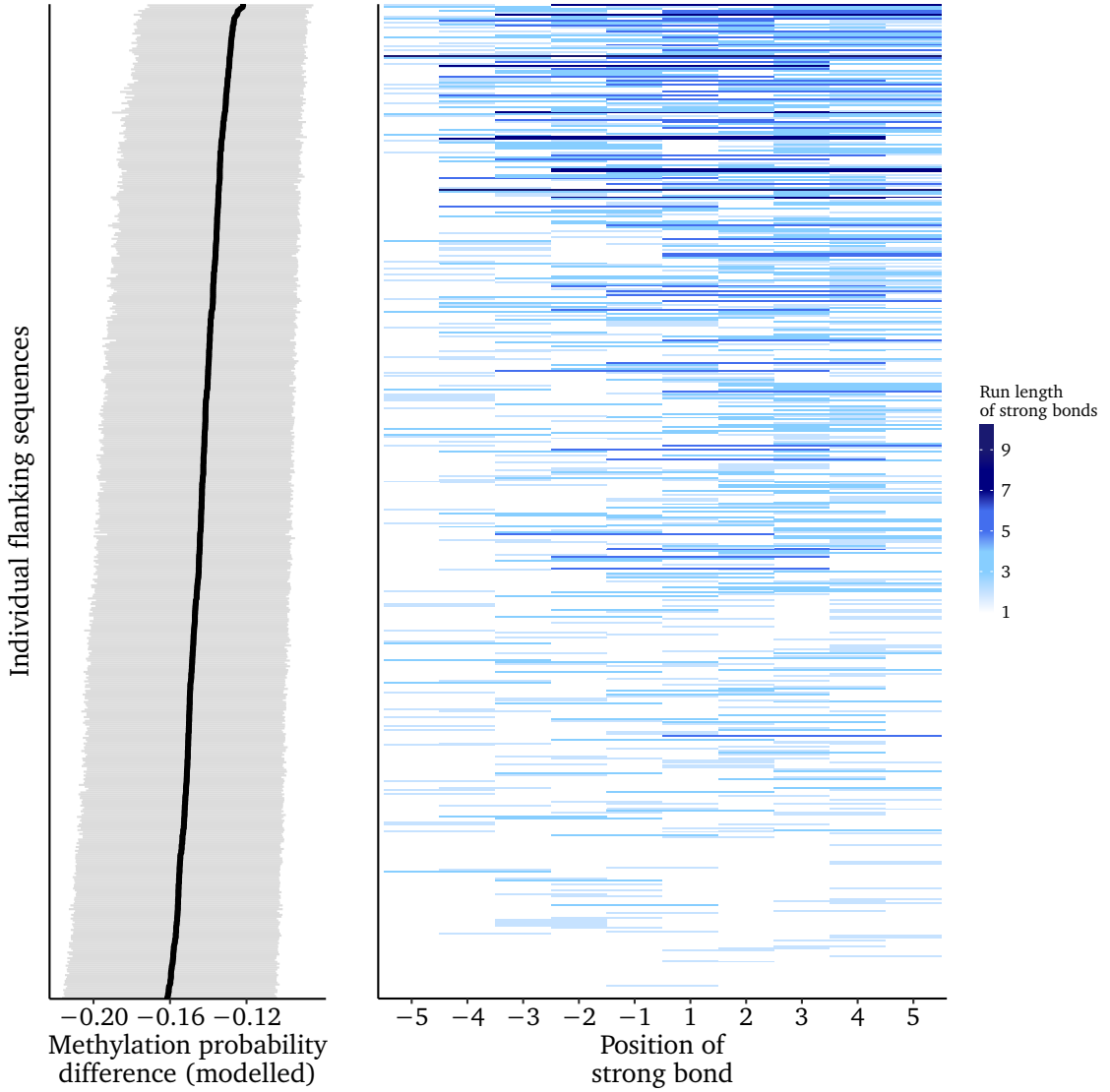


Figure S5.5: Association of methylation persistency in Dnmt1$^{-/chip}$ leukemia with the sequence directly flanking the respective CpG. On the Y-axes, individual sequences are ordered by decreasing methylation persistency from top to bottom. The left panel is reminiscent of the top panel from the previous Figure S5.4, showing the sequence's median (black) methylation persistency and IQR (gray) on the x-axis. The right panel uses the same order and items on the Y-axis, but depicts the position and consecutiveness of strong bonds in the respective sequence. The intensified blue shading in the top right corner of the heatmap indicates a slight preference for strong bonds downstream of highly persistent CpGs.

ascertained that the DNA sequence or more specifically its GC content was correlated with the degree of demethylation (Pearson's $r = 0.94$, $p < 2 \times 10^{-16}$ for 10 bp). The more CG-rich a sequence was, the lower was its tendency to demethylate [▷ Figure S5.4].

Furthermore we explored, if the exact position relative to the modeled CpG mattered. A slightly higher persistency was conferred, if the strong bonds were located downstream of the center [▷ Figure S5.5, right panel]. On the other hand the consecutiveness of the CG pairs was irrelevant, as longer runs of strong bonds did not facilitate methylation

persistency at least within the 10 bp range [▷ Figure S5.5, right panel & data not shown].

We performed many of the aforementioned analyses also for larger sections in addition to the 10 bp range, but the growing number of distinct sequences (up to 4.9×10^6 for 80 bp windows) hindered a reasonable analysis. Because many sequences occurred only once and individual aberrant sites had a great impact on the ranks, the correlation between the number of strong bonds and the methylation persistency dropped rapidly for larger windows. For example Pearson's r declined to 0.17 at a window size of 80 bp .

To circumvent some of the issues with increased windows, we resorted to a more coarse categorization to investigate the influence, which the sequence composition exerts on the methylation persistency at a larger scale. For this purpose, we resorted to the so called isochores [235], DNA segments of roughly 300 kb or less, which are distinguished by their GC content. The name still testifies that the isochores were once introduced in the context of a thermodynamic theory for the evolutionary development of genomes [236], which is nowadays obsolete [237–241]. Nevertheless, the isochore assignments relate to more recent units of genomic segmentation like LADs and TADs [242] and proved to be helpful for an approximation, to which extent the methylation persistency in Dnmt1$^{-/chip}$ c-Kit$^+$ leukemia was influenced by the GC content of the broader vicinity.

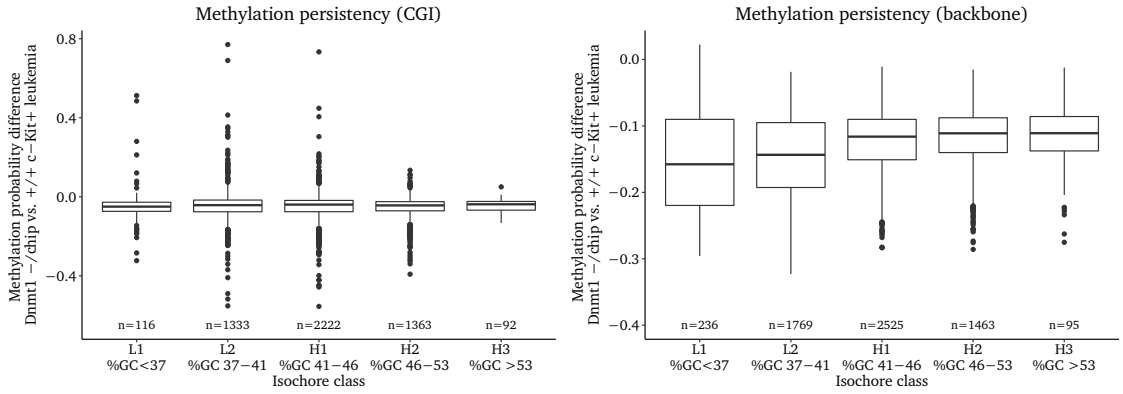


Figure S5.6: Relevance of the broader sequence composition on the modeled methylation persistency. The methylation probability difference between Dnmt1$^{-/chip}$ and Dnmt1$^{+/+}$ c-Kit$^+$ leukemia for CpG-Islands (left panel) and backbone (right panel) has been averaged over each individual isochore region and aggregated for the respective isochore categories.

We used the IsoFinder [243] tool on the murine reference genome sequence 6094 isochores in total. Then, we mapped the measured as well as modeled methylation values for CpG-Islands (CGIs) and backbone CpGs separately and averaged over each respective isochore. This ensured that a biased coverage would not affect the subsequent analysis.

As described above, CpG-Islands were mostly persistent, although a minority changed their methylation state [▷ Figure S5.6, left panel]. The altered CGIs typically underwent complete state switches and those within a particular isochore often did so in a congeneric manner. For CGIs, the surrounding sequence composition had no effect on

the methylation persistency, which seemed to be shaped by a combination of local factors and negative selection. In summary, these observations substantiated the involvement of a large fraction of CGIs in regulatory processes.

In contrast to the CpG-Islands, the backbone methylation persistency was largely dependent on the general CG-content of the DNA segment and decreased for AT-rich parts [▷ Figure S 5.6, right panel]. This was in concordance with previously published observations that hypomethylated blocks / PMDs in cancer are associated with AT-rich lamina-associated domains (LADs) [244, 245].

5.2.2 Chromosomal Insulation and Interaction

In 2016 the group of Berthold Göttgens published a comprehensive study with transcriptomic and Hi-C chromatin interaction data generated in the HPC-7 murine blood stem/progenitor cell model [246]. The cell line was used to generate comprehensive binding profiles of key transcription factors [247, 248] and a derivative in-silico model faithfully recapitulates early hematopoiesis as well as its pertubation by leukemogenic TF fusion proteins [249]. Thus, the Hi-C data of this cell line seemed promising and applicable our MLL-AF9 c-Kit$^+$ leukemia model. We considered the interactivity measured by Hi-C to be a proxy of the openness of the genomic regions, as heterochromatic areas are condensed and typically do not interact dynamically with other chromosomal regions, however they do aggregate with other heterochromatin.

We downloaded the aligned reads from the ARRAY EXPRESS repository and proceeded with the software HOMER for analysis: We built tag directories, ran quality control checks and created models of background noise. Ultimately, interaction matrices containing normalized counts at 10 kb resolution were generated, which served as input to TADTOOL [250] for calculating the insulation index/score [251].

The insulation index was computed by TADTOOL using a rectangular sliding window. On the grounds of the dataset quality, TADTOOL determined 102 353 bp as the ideal window size and used this resolution to sum up contacts within a given region. Like our modeled methylation probability, the resulting insulation score was continuous, but a cutoff could be used to call topologically associating domains (TADs). The TADTOOL authors recommended to determine this cutoff by expert guidance based on random sampling of several regions per chromosome, but given a poor repeatability (repeated random sampling of three regions on the same chromosome yielded quite deviant results) we chose a more methodological approach. We calculated the density estimates of the insulation score per chromosome [▷ Figure S5.7] and determined the local extrema of the density estimates. Most distributions comprised two local maxima and one local minimum in their center part and were only slightly shifted relative to each other (with the notable exception of the X chromosome, likely due to X inactivation). Therefore, we decided to calculate the density across all chromosomes and to use the local minimum of that function (-0.0573) as cutoff for TAD calling. [▷ Figure S5.8, left panel].

We asked, how well the insulation score and hence the interactivity of the respective

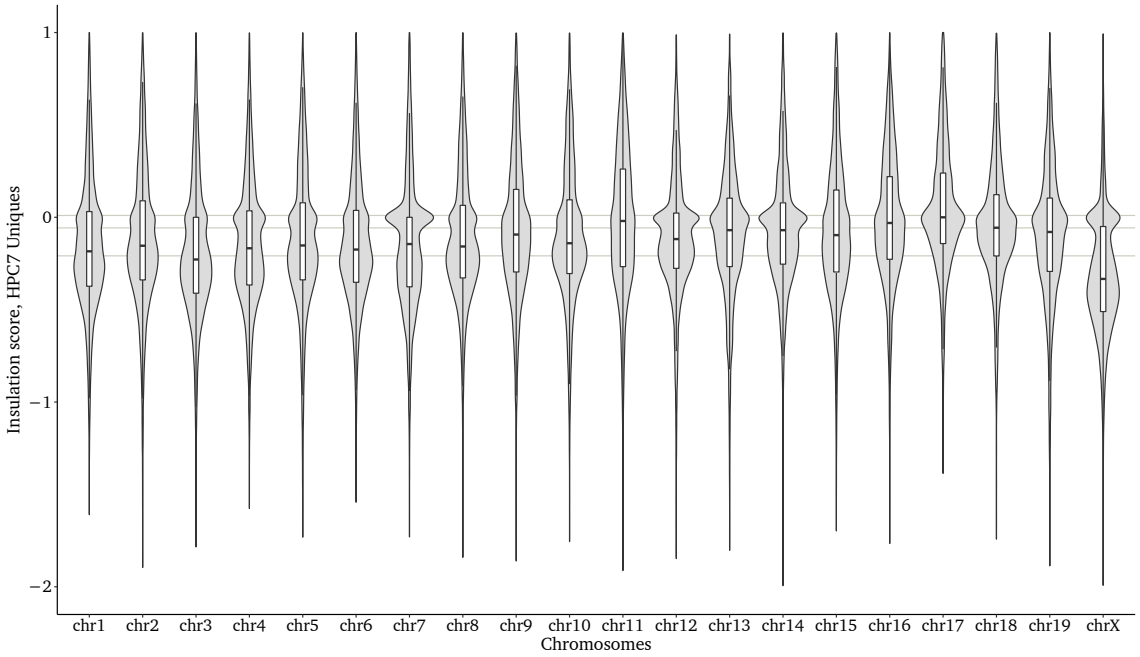


Figure S5.7: Density plots for the HPC-7 insulation score calculated for each chromosome separately. Three horizontal lines indicate the position of the local extrema of the global density estimate shown in the left panel of Figure S5.8.

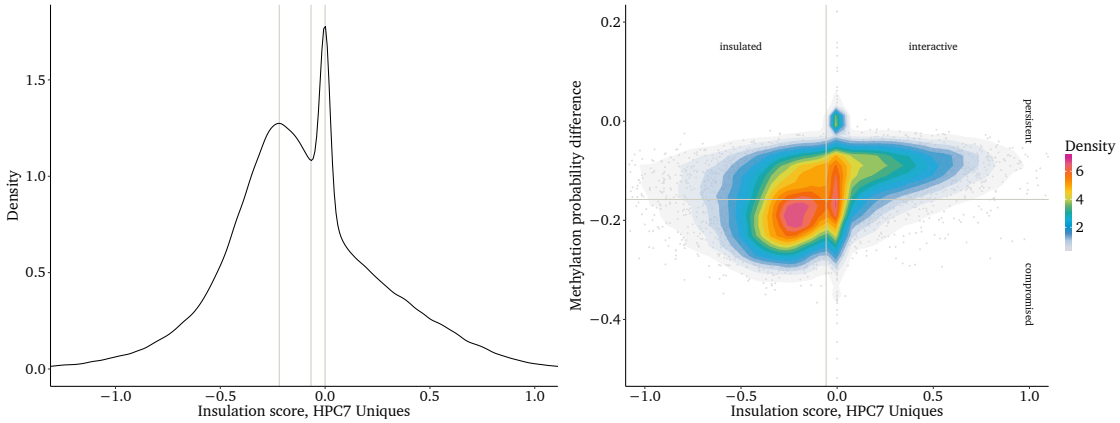


Figure S5.8: Density plots of the insulation score determined in the HPC-7 stem/progenitor cell model. The left panel depicts the global density and vertical lines mark the extrema. Two local maxima ($-0.2087, 0.0099$) and one local minimum (-0.0573) were identified and the latter used as cutoff for TAD calling. The right panel depicts the 2D density of the insulation score vs. the GAM-modeled methylation probability difference between Dnmt1$^{-/chip}$ and Dnmt1$^{+/+}$ c-Kit^{+} leukemic cells averaged over the 10 kb windows of the Hi-C dataset. The lines indicate the respective cutoffs used for categorical analysis.

genomic regions would explain the methylation persistency. Therefore, we averaged the methylation persistency difference over 10 kb windows to match it to the resolution of the Hi-C data and calculated the correlation. Globally the methylation persistency and insulation score correlated weakly, but nevertheless significantly (Pearson's product-moment correlation $cor = 0.242, t = 127.91, df = 263390, p - value < 2.2 \times 10^{-16}$). Despite the mathematical significance, the reciprocal predictability was low given the poor correlation.

At large, compromised regions typically were also insulating, while such that we found to be interacting were very rare. On the other hand among the persistent regions we observed a wide range of possible insulation scores (from very insulating to very interactive) [▷ Figure S5.8, right panel]. This was not surprising, as housekeeping genes or super-enhancers are known to possess insulating function, too [↪ subsection 1.1.1, p.6]. An example of such an insulating housekeeping gene was the highly expressed catalytic subunit of the phosphatidylinositol 3-kinase (Pik3c3), which exhibited a demethylated promoter, but otherwise resided in a methylation persistent region on chromosome 18. A small separate population was formed by CpG-Island rich areas, which exhibited intermediate insulation scores, but were highly persistent [▷ Figure S5.8, right panel].

Supplementary Chapter 6

Methylome analysis of matched non-malignant hematopoietic progenitors

6.2 Leukemia-related demethylation revisited

The new data permitted us to revisit the extent of demethylation upon transformation by MLL-AF9. For most of the project we were reliant on third-party methylome data of HSCs, which was however generated from C57BL/6 mice [215]. As the mouse strain may have significant impact on experimental results, it was important to corroborate the previous findings with matched controls from the same background.

The comparison of the new data [▷ Figure S6.1, top row] with the initial plot affirmed the previous findings in general. The vast majority of the genome was slightly hypomethylated in leukemia versus all healthy HSPC populations. The plots of both genotypes were characterized by just one density peak [▷ Figure S 6.1], which argued for a homogeneous demethylation in conjunction with leukemic transformation and supported the coexistence of persistent and compromised regions in the healthy hematopoiesis in Dnmt1$^{-/chip}$.

Nevertheless we also detected some differences. Firstly the HSCs of Dnmt1$^{+/+}$ 129/SvJae are hypomethylated by approximately 10 % in comparison to those from C57BL/6. Thus, we may have overestimated the methylation loss, which accompanied leukemic transformation by MLL-AF9. Another difference involved a set of lowly methylated regions in leukemia (25 % to 45 % methylation), which was 80 % methylated solely in Dnmt1$^{+/+}$ healthy controls [▷ Figure S6.1, left column] and shall be discussed in the next section in greater detail.

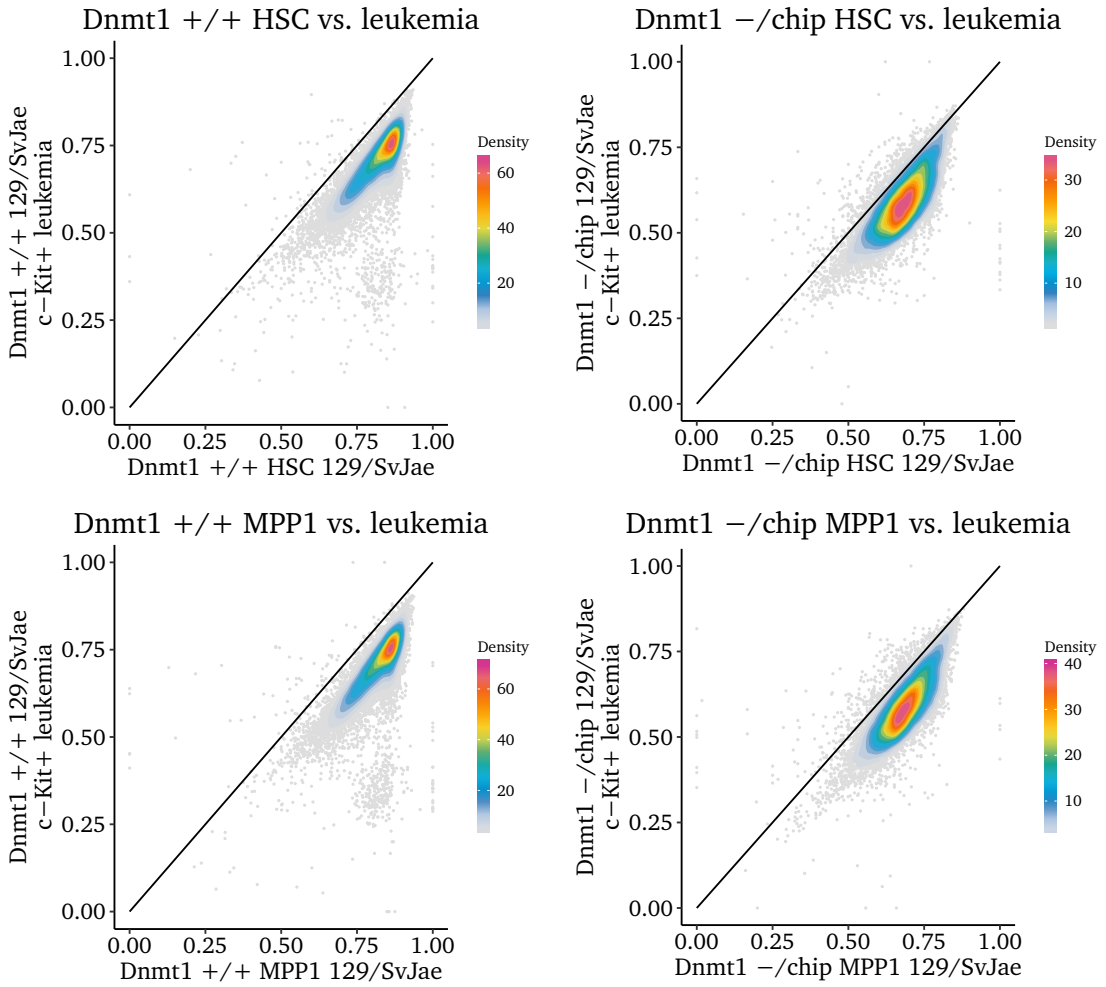


Figure S6.1: Contrast plots depicting the methylscore averages within 100 kb windows (slid by 25 kb steps) of either wild-type (left panels) or Dnmt1-hypomorphic (right panels) MLL-AF9 c-Kit$^+$ leukemia versus the genotype and strain matched healthy controls HSC or MPP1 respectively.

6.3 Compromised region addendum

In the previous chapters we presented an approach based on generalized additive models (GAMs), which permitted to locate and characterize the compromised regions in Dnmt1$^{-/chip}$ with an unprecedented fidelity. However, this method relied on single CpG methylscores, which we did not obtain as part of the prepublication access, which Daniel Lipka granted us on his laboratories' TWGBS data. Nevertheless we aimed for a better understanding of particularly those regions in healthy HSPCs and therefore approximated the compromised regions based on the 100 kb sliding window mappings. A region was considered to be compromised, if the average methylscore in Dnmt1$^{-/chip}$ c-Kit$^+$ MLL-AF9 leukemic cells $x_{c/chip}$ was less then or equal to the value $f(x_{wt})$ calculated according to Equation 6.1 [▷ Figure S6.2].

$$\text{Compromised region:} \quad x_{c/chip} \leq f(x) = 0.7 \cdot \sin((x_{wt} + 0.26)^5) \tag{6.1}$$

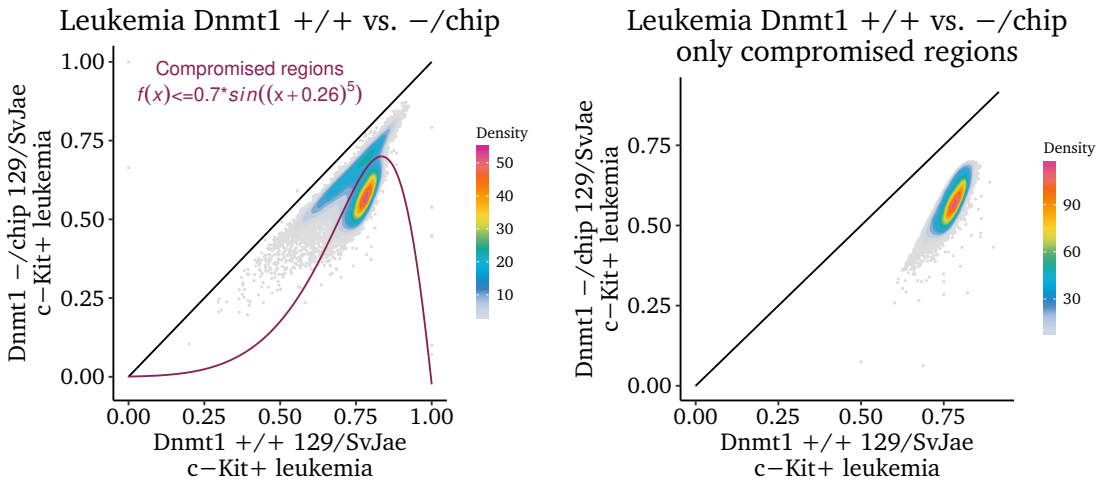


Figure S6.2: The genome was partitioned into compromised and persistent sections 100 kb in size on the grounds of Equation 6.1 and the intraleukemic comparison of MLL-AF9 c-Kit$^+$ cells.

When we plotted the compromised regions as comparisons of leukemia versus the HSCs of the concordant genotype, a minor separate population became apparent solely in the wild-type plot. The aforementioned set [↪ section 6.2] of regions demethylated from roughly 80 % to 35 % methylation upon leukemic transformation of the wild-type, while it exhibited stable intermediate methylation in any Dnmt1$^{-/chip}$ population [▷ Figure S 6.3]. Therefore, this set could be regarded as commonly compromised in leukemia of any genotype. Like the identification of a bimodal methylation pattern in healthy HSCs of Dnmt1$^{-/chip}$, also the presence of compromised regions in Dnmt1$^{+/+}$ leukemia was in disagreement with the results of the β-galactosidase (β-gal) senescence staining.

As kind of back-testing, we also created plots for the genotype contrast of HSC and MPP1 populations, which were restricted to the compromised parts of the genome [▷ Figure S 6.4]. This analysis could confirm that the compromised regions in leukemia were identical to the compromised areas of healthy populations.

Taken together, it became evident that hardly any severe perturbation such as cell cycle exit was linked to the methylation level of compromised regions. Instead, it seemed plausible that the rather arbitrary level of methylation was permitted by a lack of negative selection.

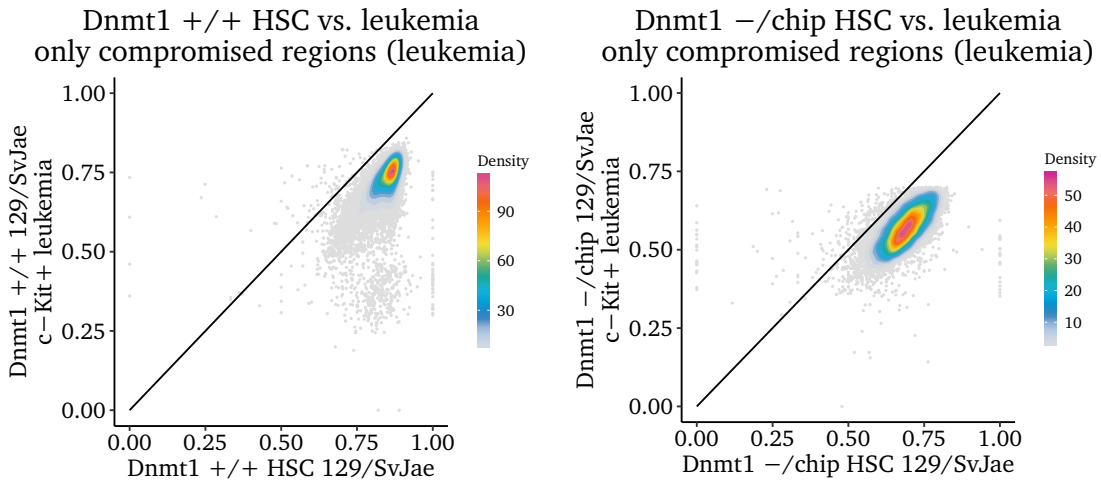


Figure S6.3: Comparison of MLL-AF9 leukemic cells with the matched hematopoietic stem cells (HSCs). The methylscore average within 100 kb sliding windows has been calculated solely for the compromised regions of the intraleukemic contrast.

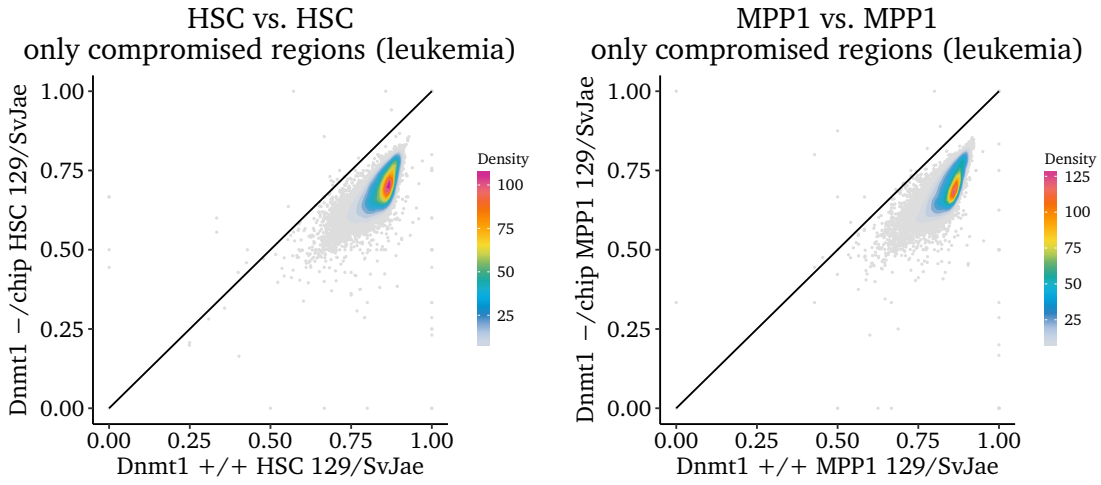


Figure S6.4: Pairwise comparison of the healthy HSPC populations of Dnmt1$^{-/chip}$ and Dnmt1$^{+/+}$ mice.

Supplementary Chapter 7

Transcriptional analysis

Contents

7.1 Characterization of Dnmt1-hypomorphic transcription

Several methods exist for the genome-wide assessment of transcription and have specific capacities, advantages and downsides. The first genome-wide dataset, which had been generated more than a decade ago in the Rosenbauer laboratory from Dnmt1$^{-/chip}$ mice comprised samples from healthy hematopoietic stem and progenitor cells [211,252] measured by microarray. The chip design "Affymetrix GeneChip Mouse Genome 430 A 2.0" represented approximately 14,000 well-characterized mouse genes. Later Lena Vockentanz, a previous PhD student, had also generated microarray data from leukemic c-Kit^{low} and c-Kit^{high} MLL-AF9 cells , but unfortunately a more comprehensive, but different chip design was chosen (Affymetrix GeneChip Mouse Gene 2.0 ST Array), which subsequently hindered straightforward comparisons to the healthy controls. Although we addressed and solved this issue by matching equivalent probesets and performing intricate normalization, those results shall not be part of this thesis.

In collaboration with Gangcai Xie from the group of Wei Chen, Lena Vockentanz had generated RNA-seq data from MLL-AF9 c-Kit^{high} and c-Kit^{low} fractions [213,214] as well as a H3K4me3 ChIP-seq from c-Kit^{high} cells for both genotypes (Dnmt1$^{-/chip}$, Dnmt1$^{+/+}$). In-depth analysis of this data was another task, which was addressed as part of this thesis work. Given that the sensitivity of RNA-seq was higher and we were intrigued by the potential impact of DNA hypomethylation on splicing, we focused on those more informative datasets [↪ section 8.2, p.57].

Indeed, widespread aberrant alternative splicing across tumors was documented by other researchers [253] at the time of writing this thesis, the cause of which however was still

disputed. Loss of methylation in gene bodies seemed to be a plausible mechanism to regulate splicing [254], however this was refuted by in an experimental in-vitro transcription model [255]. An alternative explanation was provided by the discovery of epigenetically repressed cryptic promoters, which became activated upon hypomethylation and seriously perturbed regular splicing [256]. These promoters were referred to as *treatment-induced non-annotated transcription start sites* (TINATs), because they were identified in the context of therapeutic treatment with DNA methyltransferase inhibitors (DNMTi) for cancer therapy. Previously the therapeutic effect of hypomethylating drugs for cancer therapy had been widely attributed to the reversal of focal hypermethylations silencing tumour-suppressor genes [257], but the derepression of cryptic promoters represented an appealing different mechanism.

While reviewing the aligned RNA-seq data, we had observed that a surprisingly large number of genes seemed to be incompletely annotated in the reference genome, since the canonical transcripts could sometimes not explain the distribution of sequencing reads. Often, we could spot peaks of aligned reads in the vicinity of a known gene, despite no reference exon was annotated at those loci. Furthermore the readcounts measured at the exons of a particular transcript were sometimes unbalanced, arguing against a full-length transcription of the transcript in question.

For selected genes like Pard6b, Irina Savelyeva therefore performed 5'-RACE-PCR to identify unannotated transcription start sites, which were often located inside introns and produced shorter than normal transcripts. Although we had achieved the DNA hypomethylation by means of genic Dnmt1 reduction instead of inhibitor treatment, we had thereby discovered transcription originating from cryptic promoters similar to TINATs independently from the group of Christoph Plass [256].

However, we did not consider these findings as a specific class of undiscovered transcription start sites, but rather as expectable shortcoming of the reference transcriptome, which obviously could not meet the requirements of all cell types and conditions. Even comprehensive projects like the FANTOM consortium could not provide a complete reference of all transcriptional activity in the genome, yet elaborated on the incompleteness of the existing references [10, 258].

Since fusion proteins of MLL (MLL-FP) are known to impact crucial regulators of the transcriptional machinery [259], we presumed to detect new transcripts and splice variants. To generate an experimentally determined transcriptome of MLL-AF9 cells, Irina Savelyeva in collaboration with Claudia Gebhard from the laboratory of Michael Rehli performed CAGE-seq [159, 160, 162], a technique to sequence capped 5'-ends of mRNA to determine transcription start sites. We united the CAGE-seq and RNA-seq datasets to determine an experimental transcriptome [↪ section 9.1, p.65].

7.2 Genotype validation

Like the methylome data also the transcriptome sequencing libraries of both experiments were created in collaboration with partnering laboratories and processed by several scientists and students. To avoid false conclusions due to sample confusions, we strove for a validation of the sequencing results during quality control. In contrast to the methylome data, we had access to the raw reads for those experiments, which permitted for more thorough checks, such as mapping them to control regions to confirm the presumed genotype.

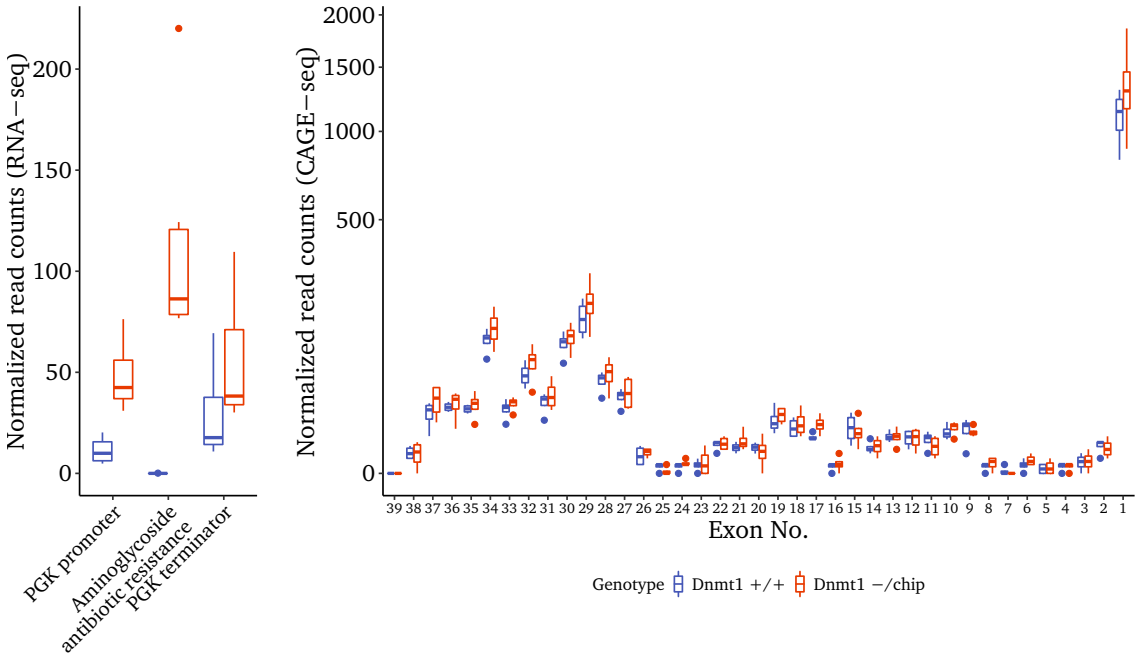


Figure S7.1: Genotype validation of the samples by remapping reads to control regions. The left panel shows the normalized RNA-seq read counts for the two genotypes (each $n = 5$), which map on the respective features. The presence of reads mapping on the resistance cassette for aminoglycoside antibiotics in the Dnmt1$^{-/chip}$ group proves the presence of the Dnmt1^{c}-allele. The right panel depicts the CAGE-seq reads allotted to the exons of Dnmt1, whose order is reversed due to their location on the antisense strand. The promoter signal is far higher than those of the exons (thus the y-axis is modulus transformed).

7.2.1 RNA-seq

Validation of the Dnmt1$^{-/chip}$ genotype in RNA-seq was relatively straightforward due to the Dnmt1^{-}-allele, which contained a PGK-NeoR-PGKpoly(A)-cassette [260, 261] excised with EcoRI/HincII from the pKJ-1 plasmid (Addgene ID: 11333). Lei et al. [262] inserted this cassette at several locations into the genomic sequence of the Dnmt1-locus to create various truncated and dysfunctional proteins. One of those constructs carries the insertion in the last third of the gene near the c-terminus and is therefore referred to as the Dnmt1^{c}-allele. The 3-half of exon 34 and the complete exon 35 located in the last third of the Dnmt1-gene had been replaced, which resulted in the protein sequence []VMAGEV-

DELETED-AGQYGV[]. Thus, almost one hundred amino acids were deleted and the tertiary structure heavily perturbed, such that the Dnmt1^{c}-allele encoded for a dysfunctional protein.

The inserted PGK-NeoR-PGKpoly(A)-cassette encompassed a functional gene copy of the enzyme neomycin phosphotransferase II (EC 2.7.1.95), which confers resistance to various aminoglycoside antibiotics, including kanamycin and G418 and originated from the transposon Tn5 [263]. Because in particular G418 also exhibits toxicity to eukaryotic cells, it can be used as selectable marker in the transformation of organisms as diverse as bacteria, yeast, plants and animals, which typically lack comparable enzymes. Therefore, we could use the sequence of the neomycin phosphotransferase II to validate the genotypes in RNA-seq: Only Dnmt1$^{-/chip}$ genomes contained the DNA cassette with the gene of the enzyme [▷ Figure S7.1, left panel]. Since expression of the neomycin phosphotransferase II was driven by the PGK-promoter, which natively belongs to the housekeeping-gene phosphoglycerate kinase 1 (PGK), the promoter and terminator sequences could also be detected in Dnmt1$^{+/+}$, albeit with a lower expression.

7.2.2 CAGE-seq

For the same reason, validation of the Dnmt1^{c}-allele in CAGE-seq was impossible, because the PGK promoter in the cassette comprised roughly 500 bp, while our single-end sequencing reads just allowed to inspect the first 35 bp of the transcript. Therefore, we could not distinguish transcription of the housekeeping-gene phosphoglycerate kinase 1 (PGK) from that of PGK-NeoR-PGKpoly(A)-cassette. While the RNA-seq results had suggested that the elevated PGK expression could suffice to set the groups apart [▷ Figure S7.1, left panel], this approach was precluded by a very high standard deviation among the biological replicates of the two groups and an inconclusive ranking in CAGE-seq expression.

While CAGE-seq is capable of highly enriching 5′-capped RNA, it is not uncommon to see weaker signals at splice acceptor sites. Thus, known exons are typically 5′-masked for downstream analyses like enhancer calling [103]. We utilized this property of the data to attempt the validation of the genotypes based on the replacement of exon 35 of Dnmt1 by the PGK-NeoR-PGKpoly(A)-cassette in the Dnmt1^{c}-allele. However, virtually no difference was detectable between the two genotype groups for exon 35 [▷ Figure S7.1, right panel].

Remarkably the signal levels varied greatly between the individual exons. For the Dnmt1$^{+/+}$ genotype this was easily explainable by alternative transcript variants and splicing. In the case of Dnmt1$^{-/chip}$, however, these reads could only arise from the Dnmt1^{c}-allele, because the Dnmt1^{chip}-allele was the spliced cDNA of Dnmt1 inserted as a minigene replacement of the genomic Dnmt1-locus [264,265]. Thus, the data suggested active transcription and post-transcriptional regulation taking place at the locus of the Dnmt1^{c}-allele despite the c-terminal truncation. However, the main goal of the analysis - the validation of the genotypes - could not be achieved for the CAGE-data.

7.3 Expression analysis

7.3.1 Stray transcription of reference transcripts

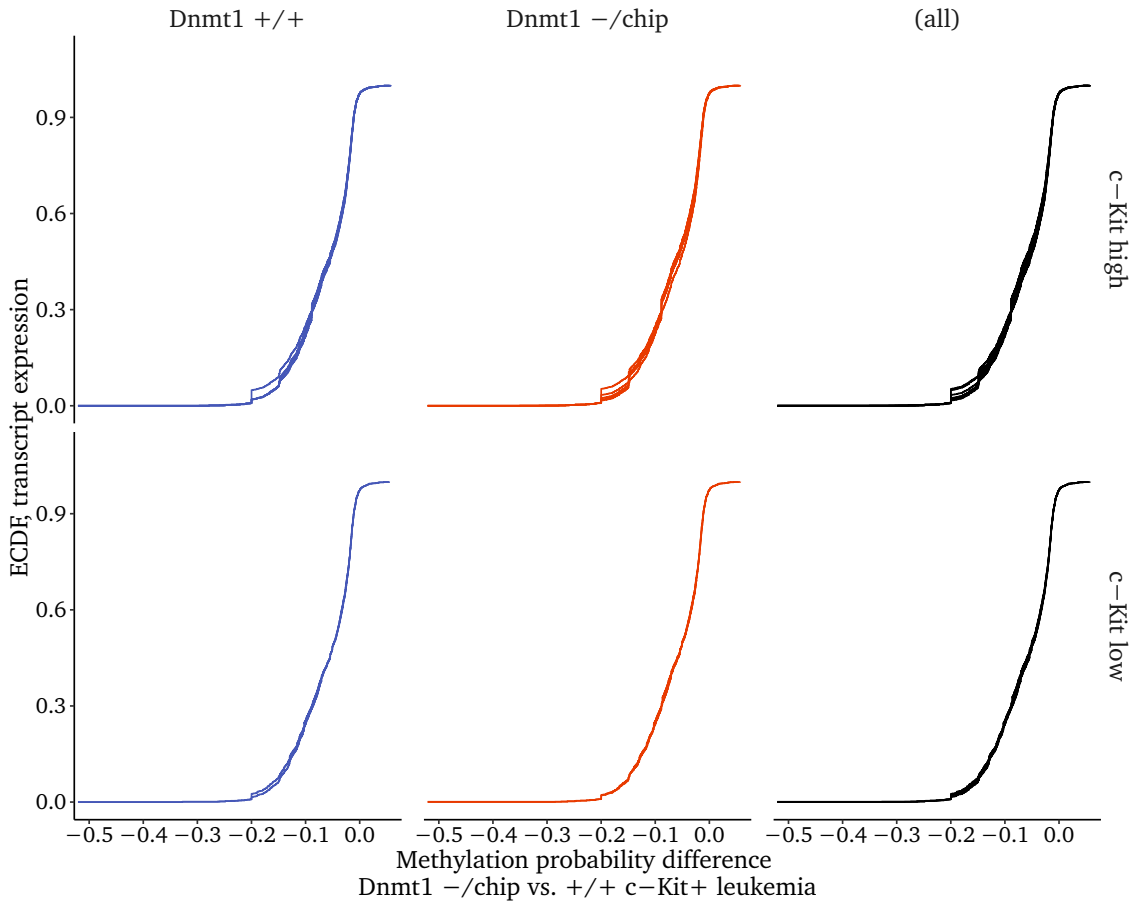


Figure S7.2: Relation of relative GAM-derived methylation probability and cumulated transcriptional expression. Each biological replicate is shown as a single curve of cumulated expression and depending on genotype and population is assigned to the respective panel. The overlay of both genotypes is shown in the last column.

As shown previously, there were few hypomethylated transcripts with a reproducibly upregulated expression across all replicates. However, this did not rule out the possibility that in single cells uncontrolled transcription due to individually hypomethylated promoters had occurred. At most, such stray transcription would have been visible in the aggregated data as increased noise.

We tried to detect and quantify possible stray transcription by several approaches. The most comprehensive analysis was based on an experimentally determined transcriptome derived from CAGE-seq data [↪ chapter 9, p.65], but we also addressed this question in the context of the reference transcripts.

On the grounds of an assumed passive demethylation in Dnmt1$^{-/chip}$ and due to the potential accessibility for the transcription machinery, we conjectured that the flexible lamina-associated domains (fLADs) might be the hot spot of stray transcription. While cLADs, which contact the nuclear lamina with high cell-to-cell consistency are gene-poor,

the fLADs harbor a notable fraction of transcripts [227]. Furthermore the variable interactions are a backbone of single-cell gene regulation and chromatin organization [266]. In connection with impaired methylation maintenance, we therefore suspected to find notable epigenetic heterogeneity and stray transcription in fLADs. However, mapping the expression of annotated transcripts onto our GAM-derived methylation persistency score resulted again in exactly comparable ECDF plots for both genotypes. We did not observe an uncontrolled onset of relevant transcription in methylation-compromised areas [▷ Figure S7.2].

We also tried to corroborate the latter result with direct chromatin data. Lacking an actual dataset of MLL-AF9 c-Kit^{+} leukemia, we resorted to Hi-C chromatin interaction data generated in the HPC-7 murine blood stem/progenitor cell model [246], which we deemed applicable to our cells [↪ subsection 5.2.2, p.39]. To derive a simple measure of the chromatin openness from the high dimensional Hi-C interaction matrices generated during our reanalysis, we applied Principal Component Analysis (PCA). Essentially, PCA introduces a new coordinate system, whose artificial dimensions are referred to as the principle components and ideally describe the data with as few dimensions as applicable. The largest possible variance in the data is covered by the first component, with the second component explaining as much of the remaining variance as possible, and so on.

When used on Hi-C data, the first principal component typically is the most informative and distinguishes *active/permissive* from *inactive/inert* compartments [38]. Therefore, we determined the value of the first principal component for all promoters, assigning them to the most likely chromatin configuration.

As expected, most of the transcribed RNA originated from areas of open chromatin and only a small fraction from more condensed areas. Since the curves of both genotypes were practically congruent [▷ Figure S7.3], we had to conclude that stray transcription of reference genes was negligible in Dnmt1$^{-/chip}$ and could confirm the previous results.

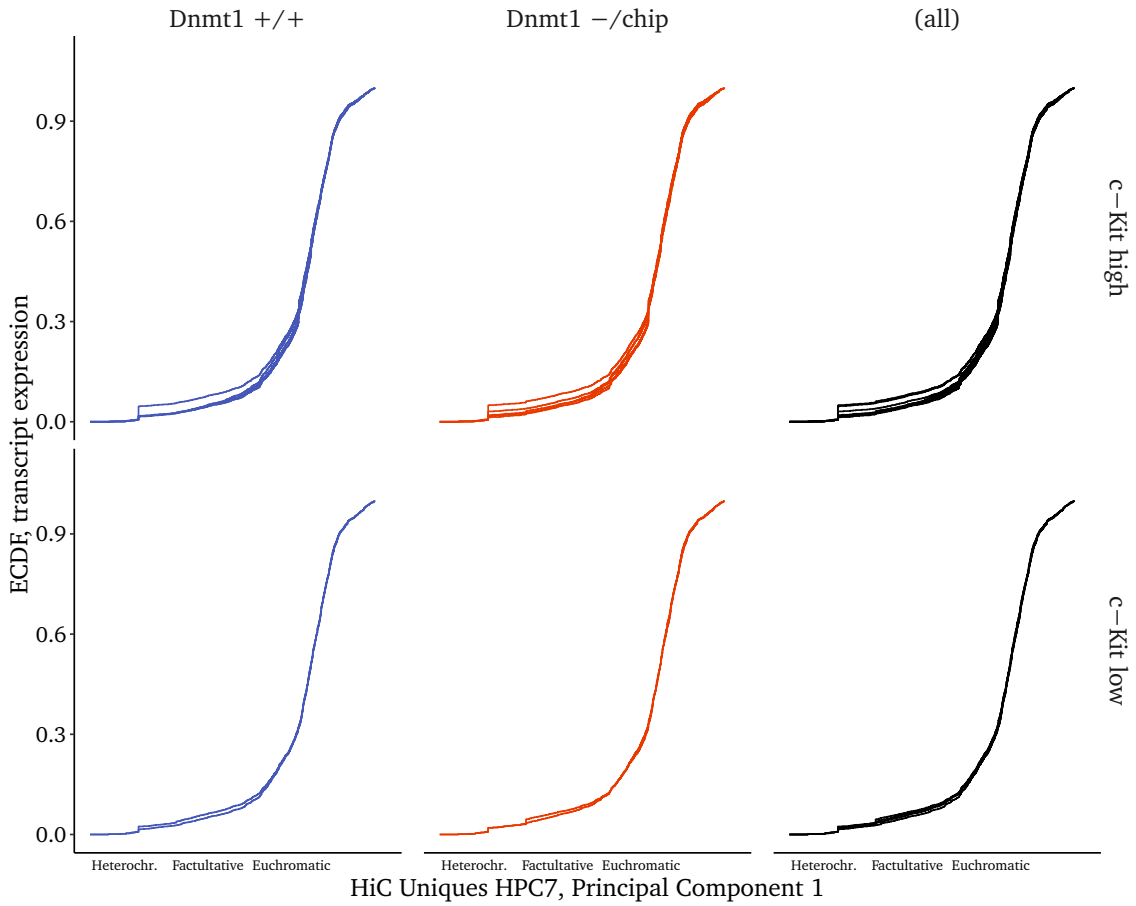


Figure S7.3: Graphs of the cumulated transcript expression relative to the first principal component computed from uniquely mapping Hi-C reads of HPC-7 cells. The latter is an approximation of the openness and interaction frequency of the local chromatin. Data for the populations c-Kit^{high} ($n=5$) and c-Kit^{low} ($n=2$) are shown as superimposed individual curves and separated by genotype.

Supplementary Chapter 8

Differentially expressed genes

8.1 Basics of differential gene expression analysis

Changes in the biological processes of a cell are typically associated with an adaption on the genetic as well as the proteomic level. Stability, post-translational modification or cellular localization of proteins may change and the transcription of previously unexpressed genes may be initiated, while that of other genes ceases.

Often, it is of interest which genes are particularly relevant for a functional alteration such as differentiation or cell cycle progression. Starting from measurements of transcript abundance in different populations or under varied conditions, such genes can be identified by means of an analysis for differential expression. The determination of genes, which are differentially expressed between two datasets, is therefore a common, yet not trivial task. Often hundreds of genetic changes can be observed in parallel owing to the complex regulatory circuitry of genomic pathways as well as the cellular heterogeneity within cell populations.

Picking suitable genes for an in-dept experimental follow-up investigation thus requires to rank genes for the best candidates. However, ranking genes solely based on test statistics (like *t*-statistics) is quite error-prone, particularly due to likely cases of misrepresentation of the population variance by the sample variance [▷ Figure S8.1]. This can be explained as follows: In presence of biological variability and experimental noise, measured values will vary. Therefore, even if we suppose an identical expression in replicate samples, measurements will give rise to a collection of deviant data values, which form a distribution. This distribution can be characterized by parameters such as its mean and variance, which quantifies how much the data points are likely to deviate from the average. Given infinitely many measurements, it would be possible to determine the distribution's parameters exactly. However, the so called unobserved true parameter values cannot be obtained in a real scenario, when only a finite number of data points can be generated. Thus, the parameters have to be estimated from the observed values.

A gene expression study will typically comprise just a few replicates for each condition, but assay thousands of genes in parallel. Although the likelihood that the observed parameters significantly deviate from the true unobserved ones is small for each gene indi-

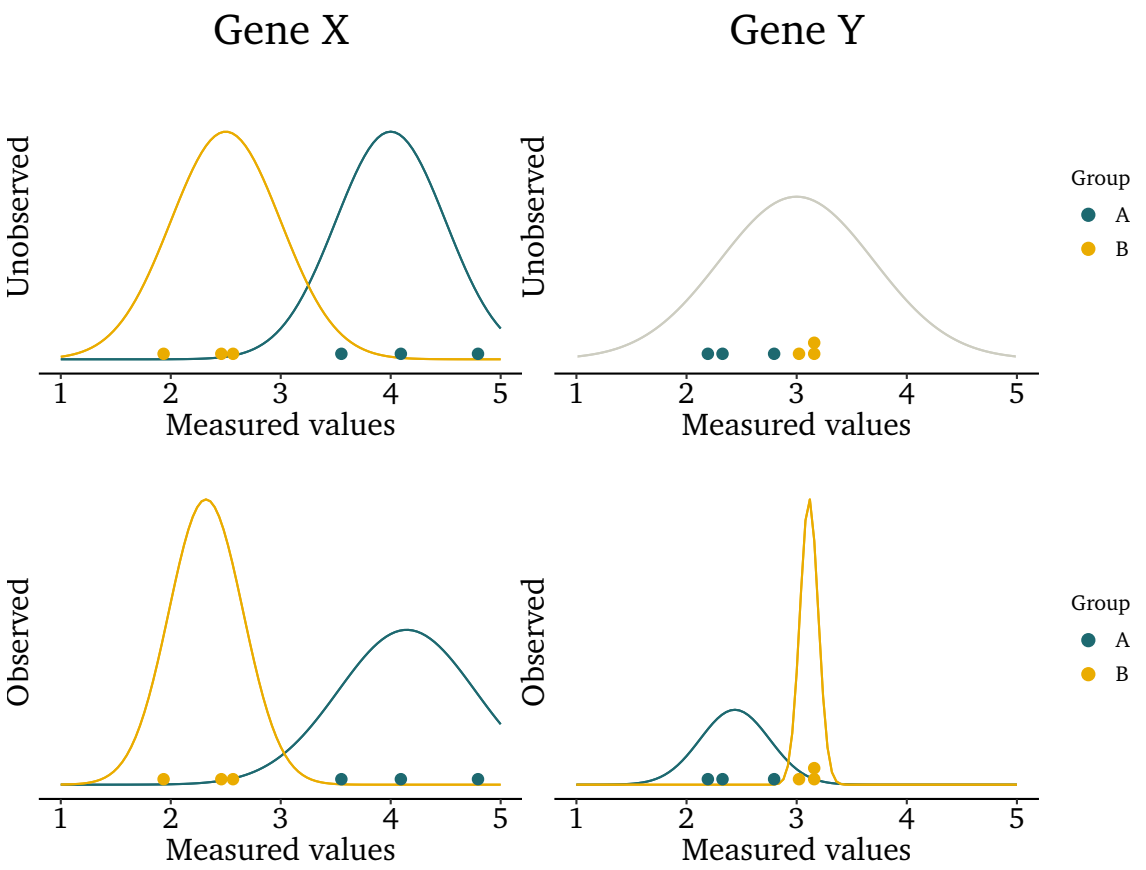


Figure S8.1: Illustration how incorrect estimation of the distributions' parameters may result in erroneous assignments. Small sample sizes per group (single values are represented by dots,3) impede precise statistical inference. In the case of gene X, the sample variance of group A was overestimated and that of group B was underestimated, while the mean values of both groups were determined approximately correctly. Nevertheless, the gene would probably be classified correctly as differentially expressed ($\mu_A = 4, \mu_B = 2.5, \text{logFC} = -0.678$). This designation would likely also be applied to gene Y, because the observed sample variance of group B happens to be unusually small by chance. Yet, all six measurements from gene Y were randomly drawn from one distribution ($\mu = 3, \sigma^2 = 0.49$), therefore gene Y is not differentially expressed.

vidually, it is probable that it occurs for some genes in the data. The most decisive parameter in regard to differential gene expression analysis is the variance. To designate a gene as differently expressed between two groups, whose true expression is confounded by experimental or biological variability, the observed expression averages must be sufficiently distinct in relation to the observed variance. Inaccurate estimation of the variance can heavily skew the test statistics and either erroneously reject or confirm the candidate gene as differentially expressed [▷ Figure S8.1]. Therefore, many algorithms and softwares for RNA-seq DE gene calling have been developed [reviewed in 267–271], which have different strengths and weaknesses and will rank genes according to various statistics. Thus, an overlap of several tools is generally not recommended. Instead a pathway analysis or gene-set enrichment may provide additional support for a specific candidate gene.

8.2 Expression changes

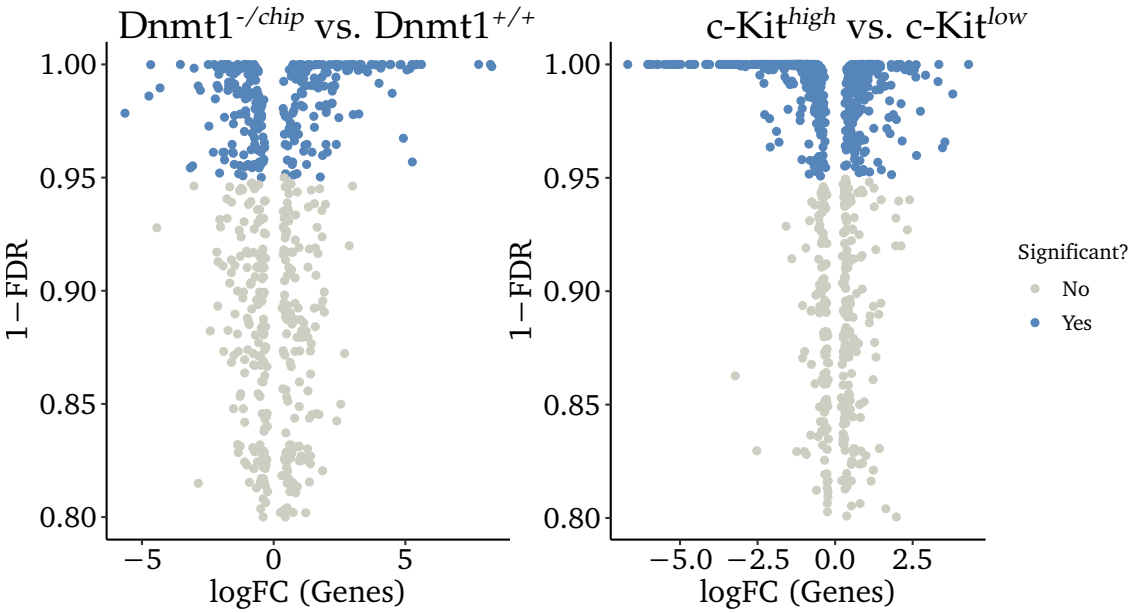


Figure S8.2: Volcano plots for both comparisons in RNA-seq, each dot represents one gene. The relative change in the expression of gene is shown log-scaled to base 2 on the X-axis, while the y-axis denotes the false discovery rate (FDR) [272] for the gene being altered. The FDR already accounts for multiple testing and the chosen significance level of 0.05 is a relatively conservative cutoff [273,274].

Lena Vockentanz in collaboration with Gangcai Xie from the group of Wei Chen had performed polyA-enriched RNA-seq from MLL-AF9 c-Kit^{high} ($n = 5$) and c-Kit^{low} ($n = 2$) fractions for each genotype (Dnmt1$^{-/chip}$, Dnmt1$^{+/+}$). We analyzed this data with the EDGER package [275] in BIOCONDUCTOR. The analysis was mostly carried out according to the RSUBREAD/EDGER pipeline [276], however we deviantly chose the *Genewise Negative Binomial Generalized Linear Models* approach [274] to test for differentially expressed genes. The alternative method applied an χ^2-approximation to the likelihood ratio statistic instead of the pipeline's default quasi-likelihood F-test, which had a more rigorous type I error rate control and was thus more conservative.

We defined two contrasts, Dnmt1$^{-/chip}$ vs. Dnmt1$^{+/+}$ and c-Kit^{high} vs. c-Kit^{low}, in the test's design matrix. Since we had sequenced more c-Kit^{high} ($n = 5$) than c-Kit^{low} ($n = 2$) samples per genotype, each group consisted of three individual ex-vivo leukemia c-Kit^{high} fractions and two paired samples of both, the c-Kit^{high} and c-Kit^{low} populations. This experimental design required to account for batch effects, therefore we used a set-up with blocking in the specification of the GLM formula.

Ultimately, we could identify 4581 differentially expressed genes (3261 individual differential transcripts) at a significance level of 0.05. For some genes, it was not possible to assign a change to a specific transcript, thus the number of differentially expressed genes surpassed that of the transcripts. Considering the respective contrasts individually, a total of 730 genes (477 transcripts) were differentially expressed in Dnmt1$^{-/chip}$ vs. Dnmt1$^{+/+}$ [▷ Figure S8.2, left panel]. In comparison, the changes for the c-Kit^{high} vs.

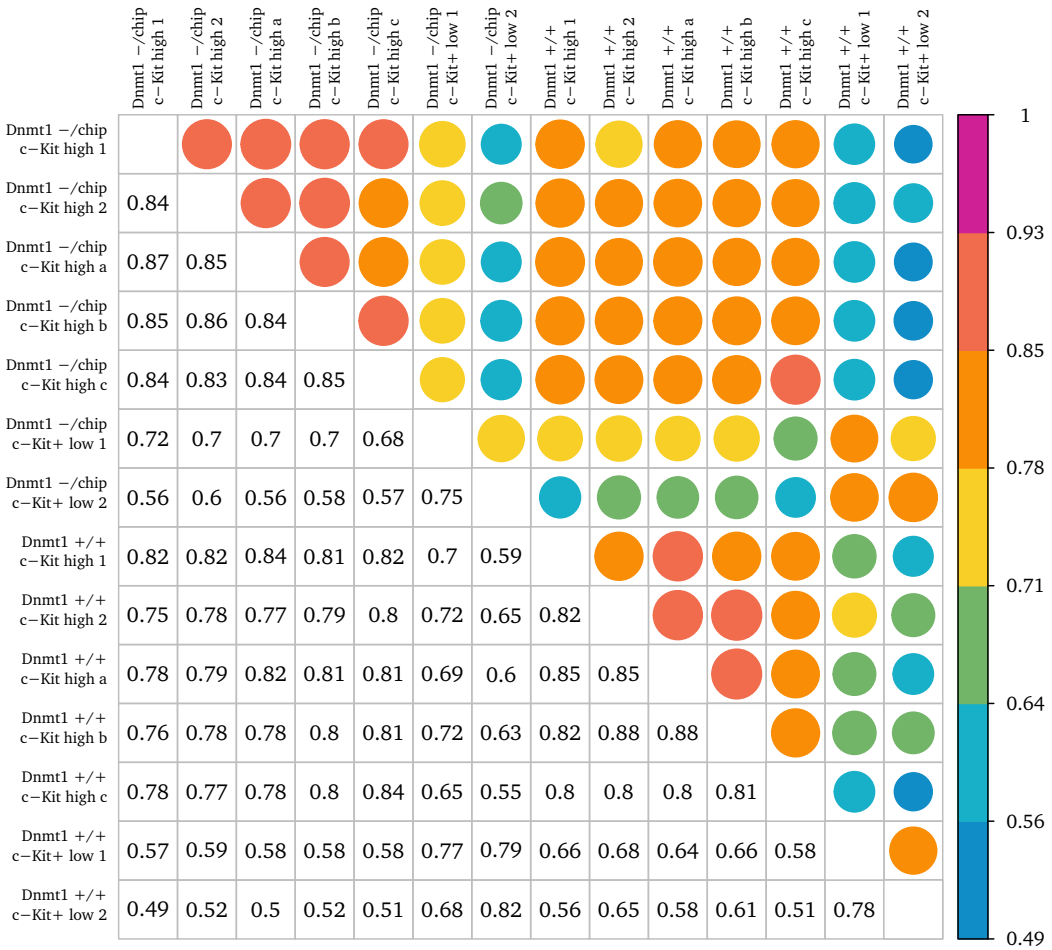


Figure S8.3: Kendall rank correlation coefficients τ of the single RNA samples. All complete observations, ie significantly differentially expressed genes for any contrast were used for the clustering ($n = 4581$).

c-Kit^{low} contrast were significantly larger as 4393 gene (3109 transcripts) differed [▷ Figure S8.2, right panel].

Evidently consequences of Dnmt1-reduction were mild compared to the distinctions between self-renewing leukemia stem cells to leukemic bulk. This was also reflected in the correlation matrix of the individual samples based on all differential genes. The correlation coefficient was increased about 0.25 for population matched samples, thus c-Kit^{low} specimen stood out clearly from the rest. In contrast, the genotype accounted for about 0.1 difference in correlation [▷ Figure S8.3] . Since the genetic basis of self-renewal and cancer cell stemness in LSCs had been addressed previously [203, 277–282] , we focused primarily on the effects of Dnmt1-insufficiency.

8.3 Contrast of Dnmt1$^{-/chip}$ vs. Dnmt1$^{+/+}$

8.3.1 Transcripts with hypomethylated promoters

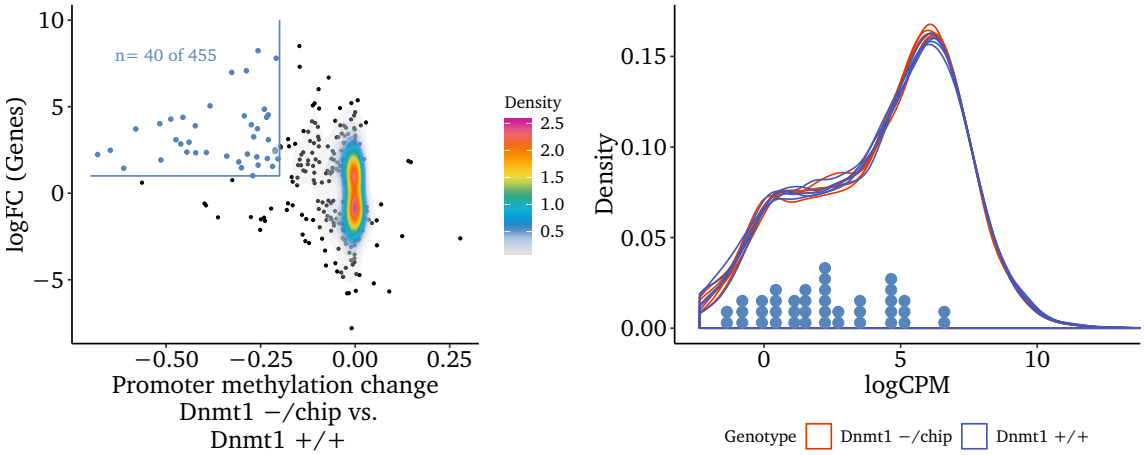


Figure S8.4: Selection criteria and expression of candidate transcripts. The left panel illustrates the selection process of the candidates listed in Table S13.3 on page 122 and puts them in relation to the remainders of significantly differentially expressed genes. The right panel shows the average expression of the candidates in relationship to those of all transcript. Every blue dot denotes a single candidate, while the colored curves represent the density of measured expression values.

The entire previous parts of section 8.3 have dealt with the differentially expressed genes and their pathways that we found when comparing the Dnmt1$^{-/chip}$ genotype with the wild-type control. Given the organization in pathways, direct regulation by promoter hypomethylation was not considered to be imperative for all differentially expressed genes. Nevertheless we mainly aimed to elucidate the ramifications of hypomethylation, thus we were hoping to identify some directly affected transcription factors.

The initial superficial analysis had already suggested that promoter hypomethylation had little relevance to overall transcription in Dnmt1$^{-/chip}$. None the less, we revisited this topic with the proper differential gene expression analysis at hand.

Our criterion for a candidate transcript was significant differential expression, an upregulation of at least 2-fold in Dnmt1$^{-/chip}$ and a hypomethylated promoter (< -0.2). All requirements were fulfilled by 40 of 455 transcripts [▷ Figure S8.4, left panel] [↪ Table S 13.3, p.122], for 22 no sufficient WGBS coverage was achieved (477 total DE transcripts). We additionally recorded 9 transcripts, which were downregulated by more than 2-fold in Dnmt1$^{-/chip}$ despite a hypomethylated promoter. The single transcript candidates were mostly expressed below average [▷ Figure S8.4, right panel]. There was no significant difference in the logCPM densities between the genotypes, which corroborated the absence of any meaningful global gene misregulation in Dnmt1$^{-/chip}$ [↪ subsection 7.3.1, p.51].

Next, we investigated whether there were matches to genes from the enriched signaling pathways, which might have produced a direct link between the hypomethylated

transcripts and the altered cellular homeostasis in Dnmt1$^{-/chip}$. However, most of the 40 transcripts were not element of the enriched pathways. Although the lysosomal Cathepsin W (Ctsw) and the endothelial Nitric oxide synthase (Nos1) were represented in the pathways, we considered them as coincidental hits due to their far downstream localization. The only candidate genes that could be linked were Interleukin-2 receptor subunit α (Il2ra) and Laminin subunit β-2 (Lamb2), which as receptor and component of the extracellular matrix, respectively, were at the very beginning of the *PI3K-Akt (mmu04151)* and *JAK-STAT (mmu04630)* signaling pathways.

Nevertheless, we focused on other candidates, notably those who would be promising new targets. Therefore, we manually reviewed the list for regulatory proteins. These included such with protein-interaction domains, DNA-binding properties or second messenger relations. Namely we experimentally tested Pleckstrin homology domain-containing, family G member 4 (Plekhg4) due to its Rho guanyl-nucleotide exchange factor activity, Protein NOV homolog (Nov) for being an activator of Notch-signaling and known essential regulator of hematopoietic stem and progenitor cell function [283] as well as the c-myc-responsive target gene modulator RING finger protein 17 (Rnf17) [284]. However, quantitative PCR in additional biological replicates of Dnmt1$^{-/chip}$ MLL-AF9 c-Kit^{high} leukemia did not confirm consistent overexpression [▷ data not shown]. The sole candidate whose expression increase was corroborated by qPCR was the ATP-dependent RNA helicase TDRD9 (Tdrd9). Since this protein mediates the repression of transposable elements (preferably during meiosis in the male germline) [285], we presumed its increased activity implied faulty transposon silencing in Dnmt1$^{-/chip}$ due to methylation loss. However, shRNA-mediated knockdown of the protein had no detectable effect on proliferation or self-renewal in MLL-AF9 leukemic clones in vitro [▷ data not shown].

8.4 H3K4me3 buffer domains

In 2014, the laboratory of Anne Brunet [286] described an epigenetic feature termed *buffer domains*. It seemed to promote transcriptional consistency and was identifiable by extended peaks of H3K4me3. This mark had previously been known to reside at transcription start sites of active genes, however they observed that it may spread far into the bodies of genes that are essential for the identity and function of the respective cell type. Since Lena Vockentanz had generated H3K4me3 ChIP-seq data from ex-vivo sorted MLL-AF9 leukemic stem cells (c-Kit^{high}), we applied the published analytical approach to characterize those buffer domains in the leukemia and elaborate on possible differences to Dnmt1$^{-/chip}$.

As described in the paper [286] we considered the 5 % broadest of all H3K4me3 peaks to be buffer domains. It should be noted that the peak caller MACS2 [287] used in the original publication was designed for transcription factor ChIP-seq data with narrow peaks and was, despite improvements and the introduction of the `--broad` option, known for a weak performance with regard to histon ChIP-seq data [288]. Therefore, we replaced it by SICER [289], which exhibited superior abilities to call clustered enriched regions

from diffuse data [290,291] and was thus better suited for broad peaks such as H3K36me3 or buffer domains of H3K4me3.

We deliberately allowed for small gaps of less than 2 kb and merged close adjacent peaks to larger domains prior to defining the buffer domains. This merger and particularly the application of SICER increased the maximum breadth from 5.8 kb to 235 kb, while the median size of a buffer domain grew to 17.5 kb. Because of these alterations, both were now in line with published reference values. Intersection of the buffer domains in Dnmt1$^{+/+}$ and Dnmt1$^{-/chip}$ ($n = 2895$ and $n = 2455$ respectively) showed that the genotype disparity in buffer domains (Jaccard index $= 0.41$) was higher than that in all H3K4me3 peaks (Jaccard index $= 0.58$). Therefore, we anticipated that the impairment of Dnmt1$^{-/chip}$ cells to acquire and maintain malignant self-renewal properties arose at least in part from the deviant buffer domains.

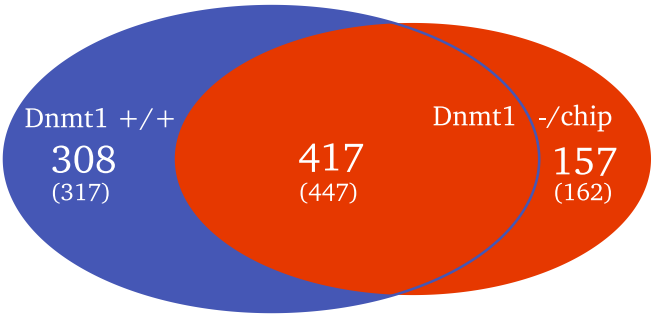


Figure S8.5: Number of reference genes, of which at least one transcript (count given in brackets below) was marked by a broad H3K4me3 domain. However, most buffer domains in either genotype did not overlap any reference transcript.

Since it was suggested that buffer domains can be used as a discovery tool to identify important regulators of cell identity [292], we aimed to elaborate on the differences associated with a Dnmt1 hypomorphic genotype and to identify the altered genes. Surprisingly, just 553 (19.1 %) of the Dnmt1$^{+/+}$ and 418 (17.0 %) of the Dnmt1$^{-/chip}$ buffer domains overlapped any annotated gene at all, thus the genesis and function of the other peaks remained elusive. Most buffer domains overlapped just a single annotated gene, however some peaks extended to multiple (up to 24) at the same loci. Broad H3K4me3 domains were neither associated with higher (normalized) ChIP intensities nor correlated with the number of transcripts or the length of overlapped genes [▷ data not shown].

A majority of buffered genes was shared among the genotypes [▷ Figure S 8.5] and seemed to exhibit an elevated median expression [▷ Figure S 8.6, rightmost panel]. To test the significance, we fitted a generalized linear model to the measured expression as response variable and used genotype together with the peak status as predictors. Decomposition of the model into its linear hypotheses and calculation of Tukey's all-pair comparison showed that the expression difference of *Regular H3K4me3 peaks vs. none* was not statistically significant [▷ Table S8.1]. In contrast to published results [286], the expression of genes marked by H3K4me3 buffer domains was significantly elevated com-

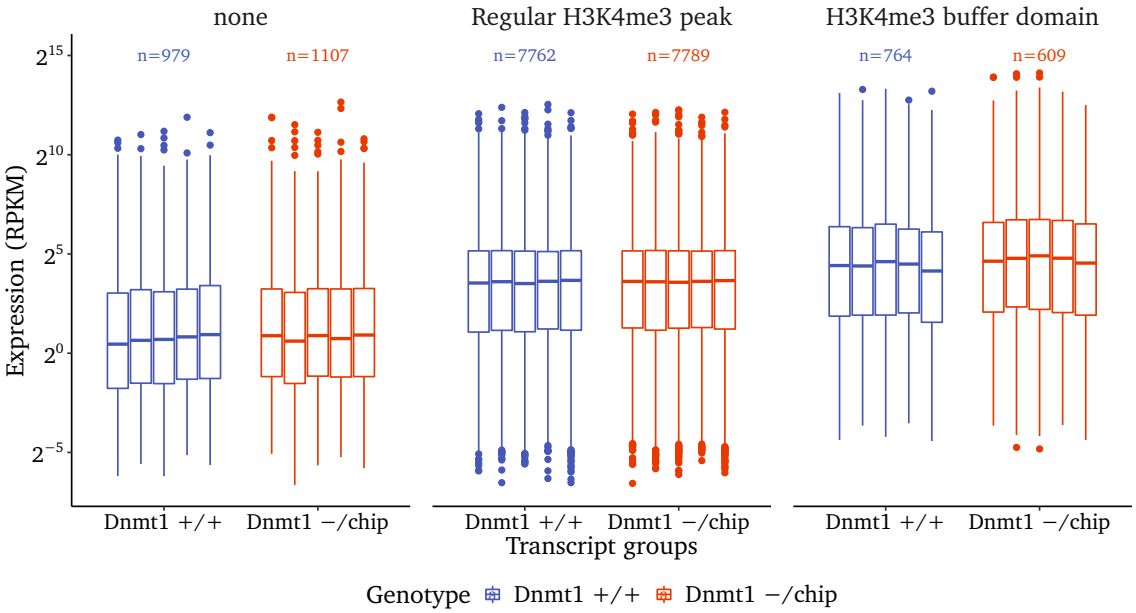


Figure S8.6: Boxplots of transcript expression depending on the H3K4me3 coverage. Transcripts are categorized based on presence of a promoter peak (−500 bp to 100 bp), full length coverage by a buffer domain or the absence of both. Measured expression of the genes in single c-Kit^{high} biological replicates is summarized and shown as boxplots. Numerical representation of the plotted median expression can be found in Table S8.1.

Genotype	H3K4me3 category	Median expression (RPKM)			
Dnmt1$^{+/+}$	none	1.62		0.277	$<1 \times 10^{-4}$
Dnmt1$^{-/chip}$	none	1.75			
Dnmt1$^{+/+}$	Regular H3K4me3 peak	12.0	$<1 \times 10^{-4}$		
Dnmt1$^{-/chip}$	Regular H3K4me3 peak	12.24		2×10^{-4}	
Dnmt1$^{+/+}$	H3K4me3 buffer domain	21.15			
Dnmt1$^{-/chip}$	H3K4me3 buffer domain	26.83			

Table S8.1: Median expression of the data shown in Figure S8.6. The brackets to the right indicate the resulting, adjusted p-values for each contrast by Tukey's all-pair comparison of linear hypotheses.

pared to both other categories. We also detected a meaningful difference for the genotype nested in the H3K4me3 buffer domain category.

Given the increased expression of buffer domain genes in Dnmt1$^{-/chip}$, we investigated a potential promoter hypomethylation. However, the promoters of any H3K4me3 marked gene were essentially methylation-free [▷ Figure S8.7, middle and right panel] and the promoters of the genotype-specific transcripts (317 and 162 respectively) [▷ Figure S 8.5] did not feature contrasting methylation rates [▷ data not shown]. Solely unmarked transcripts were frequently methylated and consequentially not expressed. The median

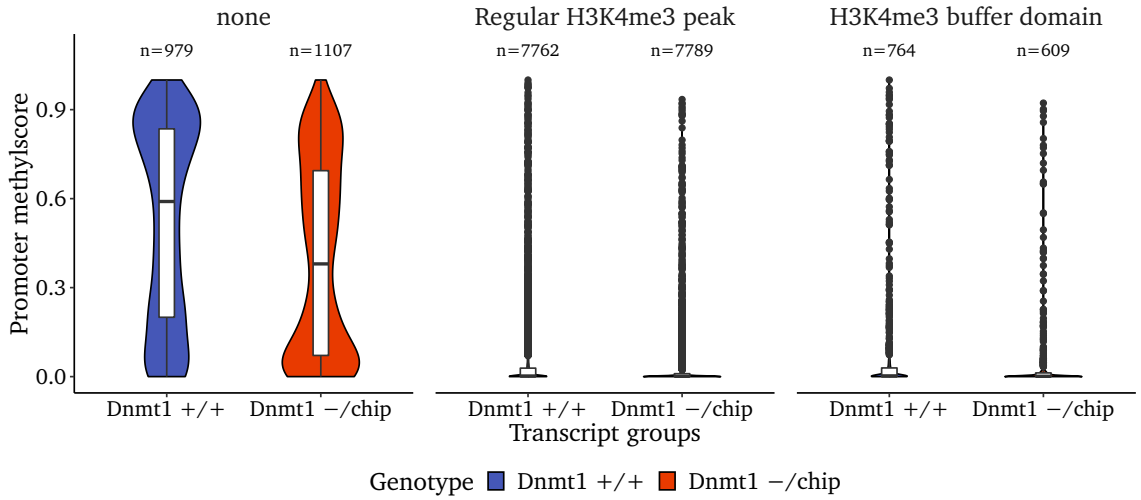


Figure S8.7: Average promoter methylation rate (−500 bp to 100 bp around the TSS) for the transcripts, which fell into the respective H3K4me3 categories.

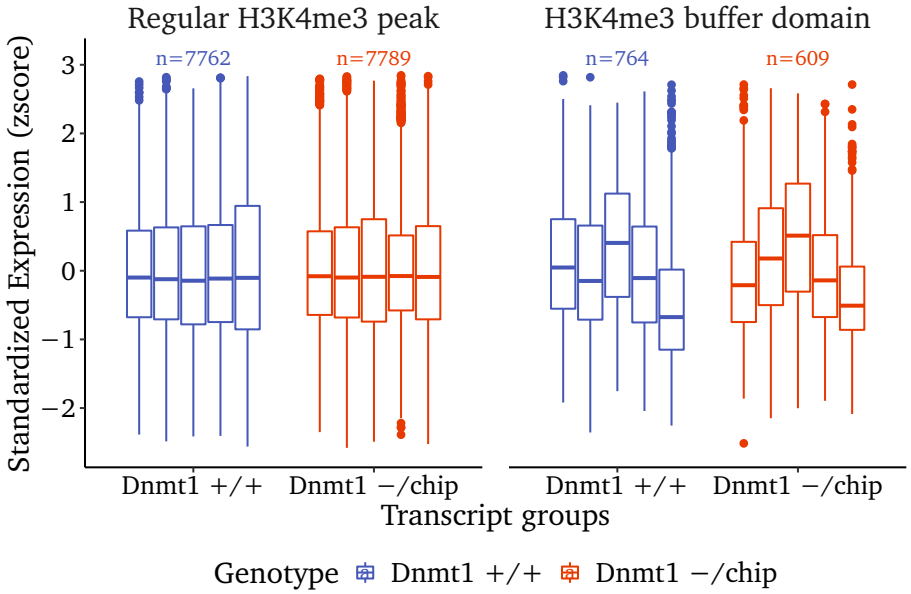


Figure S8.8: Boxplots of the standardized expression, which was calculated for each transcript individually. Raw values were centered by having the corresponding mean expression of the particular gene subtracted from them. Scaling was performed afterwards and refers to dividing the centered values by the sample standard deviation of the set. As the sum of all standardized values per transcript equals 0, purely random deviations should ultimately cancel out over hundreds of genes, which was not the case for those marked by buffer domains.

methylation rate of unmarked promoters was 0.59 and dropped notably in Dnmt1$^{-/chip}$ to 0.38 [▷ Figure S8.7, leftmost panel], which however did not translate into a significantly increased expression of the hypomethylated genes [▷ Table S8.1].

Supplementary Chapter 9

Experimental transcriptome

9.1 Assembly of non-reference transcripts

In principle, there are two approaches to establish a custom transcriptome from RNA-seq data. One can opt for a true de novo assembly [293], which combines overlapping sequencing reads into longer continuous genomic sequences (so called contigs). However, for most bioinformatic approaches to the problem (like *De Bruijn graphs*), there is a trade-off between ambiguity as well as efficiency and computational demands such as memory consumption. Thus, for common model organisms, for which reference genomes exist, it is generally preferable and more precise to align the reads first to the genome and reconstruct the transcriptome from those alignments [294–296].

In addition to RNA-seq data for the transcriptome assembly, we could fortunately also utilize CAGE-seq data to determine and validate transcription start sites [↪ section 7.1, p.47]. On the downside, the read specifications of both experiments (single-end, 36 bp short, Phred QS $\leq$ 35) no longer corresponded to the state of the art, therefore we had to adapt some of the tools' defaults. Briefly, we united the reads per genotype and aligned them to the *NCBI37/mm9* reference genome with BBMAP [297] using some custom settings[1]. Afterwards we employed STRINGTIE [298] to reconstruct the transcripts and retained only those assemblies, which were supported by a 5'-prime CAGE-seq signal with a custom script. Further own AWK scripts were applied to finalize the transcriptome reference `gtf`.

In total, we were able to reconstruct 43 597 elongated transcripts from CAGE-seq confirmed transcription start sites, but just 12 850 (29.47 %) were ultimately considered for downstream analyses. This was due to the additional requirement to be expressed in at least two samples irrespective of their genotype [↪ section 7.3, p.51] and most transcripts were either specific to one sample or artificial mergers of reads originating from different samples.

The remaining transcripts were subjected to differential expression analysis as described before [↪ section 8.2, p.57], which confirmed the previous findings [▷ Figure S8.2]: Con-

[1] `minid=0.9 padding=6 tipsearch=200 maxindel=500000 intronlen=5 ambig=random sssr=0.97`

siderably more transcripts were differentially expressed between c-Kit^{high} and c-Kit^{low} cells ($n = 3686$) than between the genotypes Dnmt1$^{-/chip}$ and Dnmt1$^{+/+}$ ($n = 519$) [▷ Figure S9.1, left panel].

Subsequently, we ran GFFCOMPARE [299] on all three gene sets to classify and categorize the transcripts. The respective status codes are listed in Table S9.1. Unsurprisingly intron-exon-matched annotated transcripts were the dominating class (**=**) in all sets (68 % to 85 %). The second most common category was composed of unspliced pre-mRNA fragments (**e**), which predominantly bridged introns at the 3′-ends of the transcripts. Usually, the last few or at least the final and penultimate introns exhibited up to 40 % of the signal strength of the adjacent exons. We considered this to be a technical artifact of the poly(A)-enrichment during library prep and a reason for the 3′-bias in the RNA-seq data. The third largest fraction was unique, intergenic (**u**) transcripts with no direct relation to annotated genes. These were mostly short (<2 kb), unspliced fragments typically framed by SINEs or LINEs, possibly pseudogenes, sequences of viral origin or transposons. Their expression or number did not increase in Dnmt1$^{-/chip}$, therefore they were probably not attributable for the impairment of self-renewal, although their exact role remained elusive.

Status code	Status explanation
=	Exact match to reference transcript
c	Contained within the reference transcript but incomplete
j	Potential novel isoform that shares at least one splice junction with a reference transcript
e	Overlaps a reference exon plus at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment
i	Falls entirely within a reference intron
k	Falls entirely within a reference intron, reverse containment (antisense)
o	Exon of predicted transcript overlaps a reference transcript
p	Lies within 2 kb of a reference transcript (possible polymerase run-on fragment)
r	Has >50% of its bases overlapping a soft-masked (repetitive) reference sequence
u	Is intergenic in comparison with known reference transcripts
x	Exon of predicted transcript overlaps reference but lies on the opposite strand
s	Intron of predicted transcript overlaps a reference intron on the opposite strand

Table S9.1: Categories and status codes assigned by GFFCOMPARE to the de novo reconstructed transcripts.

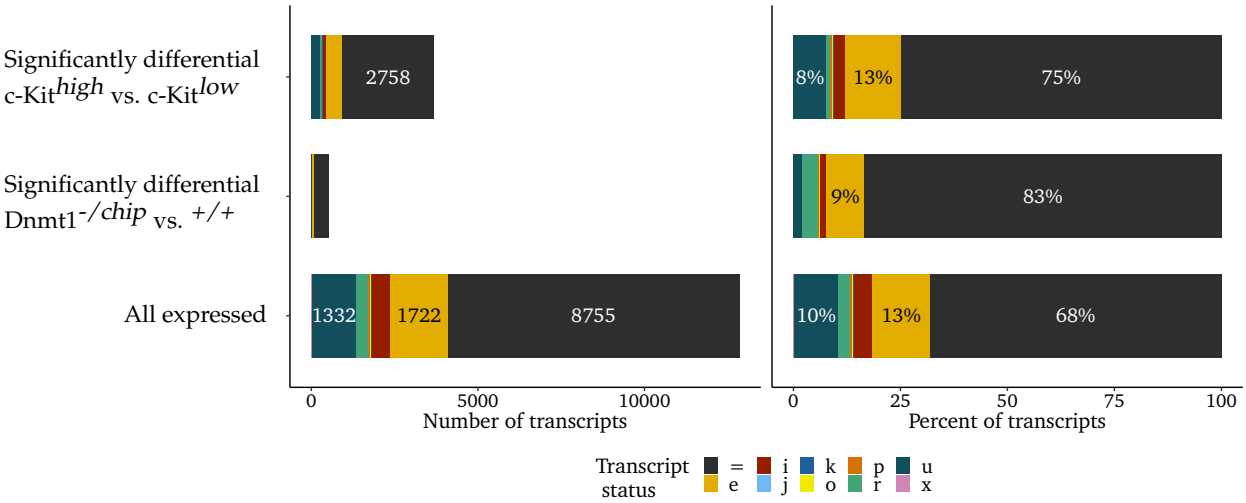


Figure S9.1: Absolute counts (left panel) and relative shares (right panel) of the respective status codes [↪ Table S9.1] within the three gene sets under investigation. The largest set comprises the 12 850 transcripts expressed in at least two samples, the others the significantly differentially expressed items for a particular contrast.

9.2 Expression of non-reference transcripts

While the counts of a transcript class may be informative in terms of gene regulation or the lack thereof, it is certainly the expression, which confers biological relevance. Therefore, we mapped the reads of each sample individually on the assembled transcripts and detected pronounced differences in expression for the classes [▷ Figure S9.2].

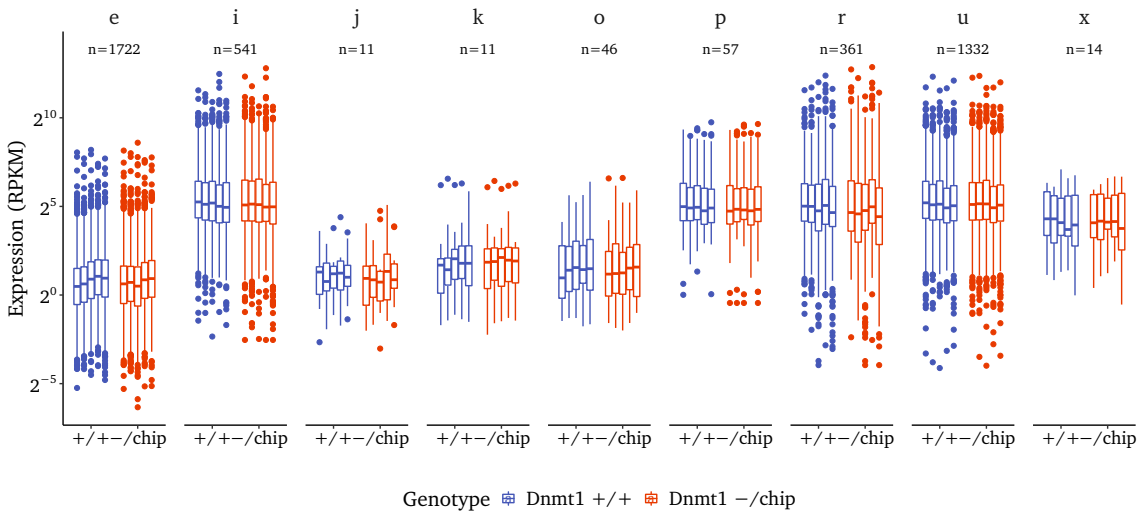


Figure S9.2: Boxplots of the all expressed, unannotated ($n = 4095$) transcripts' measured expression (by RNA-seq). Each biological replicate is shown as separate box and colors indicate the respective genotypes.

Unspliced pre-mRNA transcripts (**e**) constituted the second most common class, but their expression was obviously subject to normal genic regulation by the transcripts' regular promoters and thus comparable to that of the full length references [▷ Figure S8.6, p. 62].

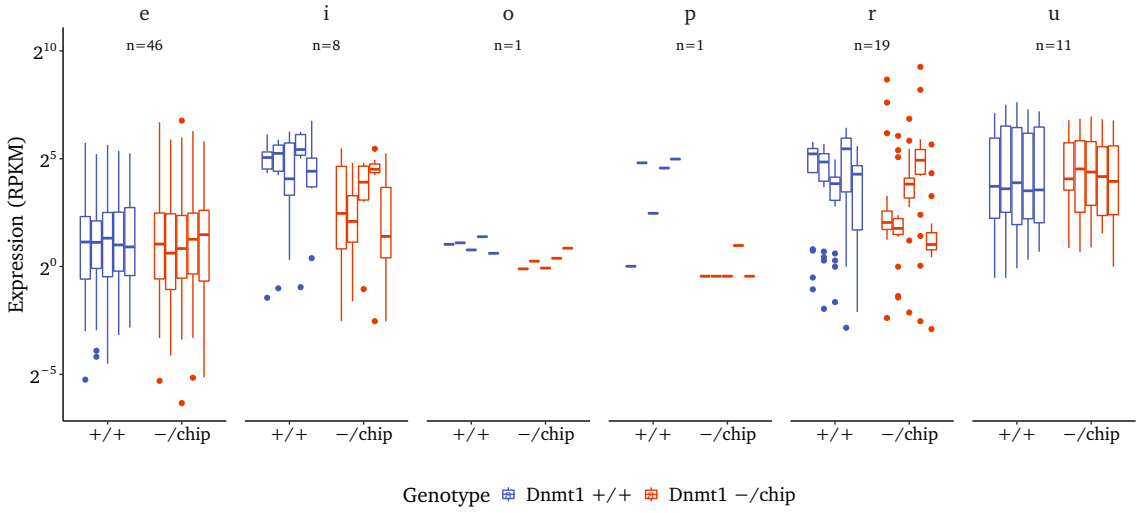


Figure S9.3: Expression of unannotated transcripts, which were differentially expressed between Dnmt1$^{-/chip}$ vs. Dnmt1$^{+/+}$ ($n = 86$). Colors represent the genotypes and each biological replicate is shown individually.

Comparably lowly expressed and relatively scarce were transcripts of the classes **j**,**k** as

well as **o**. Although those transcripts with novel splice junctions (**j**) might have played a role in the leukemogenesis, none of them was differentially expressed and thus of interest for this project. Our data therefore did not corroborate widespread joining of RNAs originating from cryptic promoters to regular reference transcripts during splicing as described before [256]. However, in our mouse model, we have achieved the DNA hypomethylation by means of genic Dnmt1 reduction instead of inhibitor treatment, which may explain the differences.

Despite just having a different directionality, intronic sense transcripts (**i**) were higher expressed and remarkably much more frequent than intronic antisense transcripts (**k**).

In terms of expression, the most eye-catching classes were the **r** and **u** transcripts, which we however considered to be functionally and biologically equivalent in general. The reason was that most transcripts classified as unique intergenic (**u**) were short (<2 kb) fragments, sharply demarcated by non-transcribed SINEs, LINEs or other repeats. Therefore, many **u** items were at least affiliated with repeats, although they did not extend into the repetitive sequence itself. However, this might have been a technical issue, since the short, single-end reads hardly permitted unique or high-confidence secondary alignments in repeats [↪ section 9.1]. Depending on the surrounding sequence, many **u** transcripts could thus have been incompletely reconstructed and just for this reason failed to meet the **r**, requirement of >50% of overlap with a repetitive reference sequence [▷ Table S9.1].

Considering that the importance of long non-coding RNAs for mouse acute myeloid leukemia development has already been described [300] and we only detected very few long unannotated transcripts, a significant fraction of incomplete assemblies across all RNA categories must be assumed, especially in regions of low mapability. Since we required the presence of 5′ CAGE-seq signals supporting transcriptional initiation and otherwise discarded an assembled transcript, the majority of incomplete transcripts should be 3′ truncated.

9.3 Methylation of non-reference transcripts

In previous chapters the effects of hypomethylation in Dnmt1$^{-/chip}$ leukemic cells on reference transcripts have been addressed in great detail [↪ section 8.3, p.59]. Most of these analyses were repeated on the basis of the experimental instead of the reference transcriptome, but without relevant different findings [▷ data not shown].

However, published literature had suggested that DNA hypomethylation is capable of reactivating cryptic, dormant promoters, which were referred to as *treatment-induced non-annotated transcription start sites* (TINATs) [256]. Therefore, we performed an in-depth follow-up analysis of the promoter region methylation rate by transcripts classes.

Regular promoters of annotated transcripts were either unmethylated or just weakly methylated [▷ Figure S9.4, upper left panel], which was in accordance with the com-

monly accepted notion that promoter methylation represses transcription. This held true even for the methylation data from hematopoietic stem cells, although the RNA for the de novo assembly originated exclusively from MLL-AF9 leukemic cells. This suggested that even leukemia-specific transcripts already exhibited an unmethylated or lowly methylated promoter in regular HSCs.

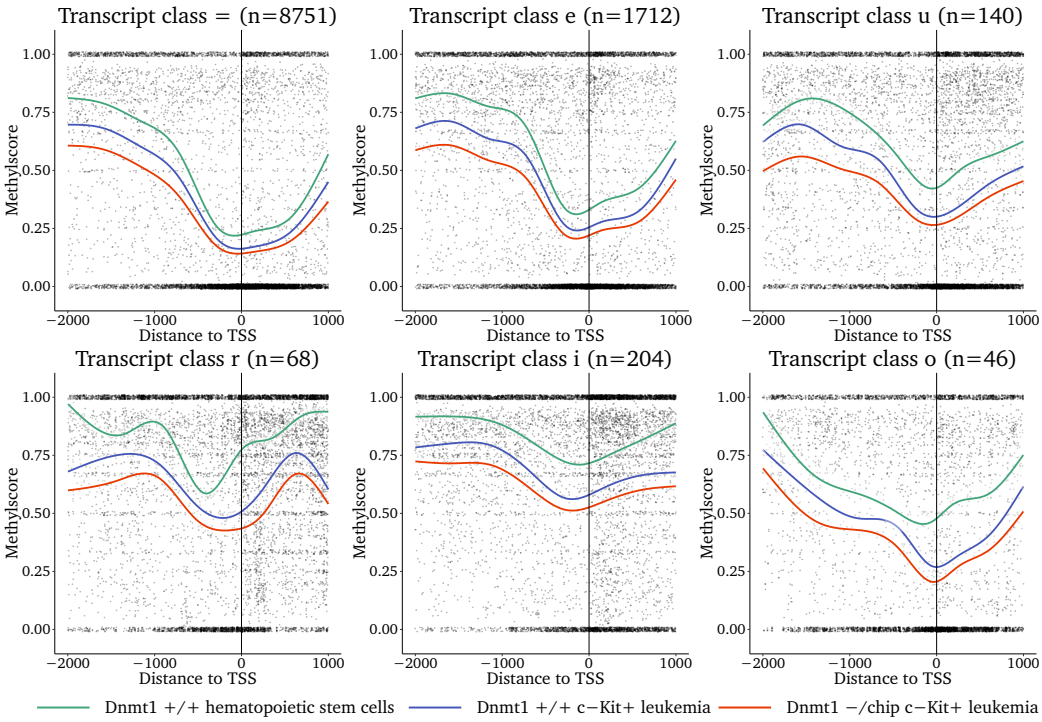


Figure S9.4: Methylation in the CAGE-seq supported 3 kb promoter region of de novo transcripts. Colored lines represent the smoothed average methylscore, which is displayed on top of the measured methylation rate of single CpGs (black dots). CpGs without sufficient WGBS coverage (3 reads) or transcripts without (covered) CpGs are not shown.

In theory, unspliced pre-mRNA (**e**) as a precursor of reference transcripts should fully map to the same promoters, however typically the assemblies just comprised the last few 3'prime exons of a transcript. It remained elusive, whether the detected 5'prime ends corresponded to the true TSS or the transcripts were incompletely assembled and one of the typical exon CAGE-seq signals was mistaken for the initiation site. In terms of methylation however, the presumed TSS of **e** transcripts were significantly hypomethylated compared to those of transcripts with an true intronic initiation (**i**) [▷ Figure S9.4, center panels].

In the latter case, the stable methylation might be explainable by persistent gene body methylation, whereas that **r** transcripts may be attributable to the need of silencing of neighboring transposons or viral genes. Lower, intermediate methylation was detected at unique **u** intergenic transcripts as well as at the genic exon-overlapping assemblies **o** [▷ Figure S9.4].

All described patterns were however not sample-specific, the methylation ranking for all transcript classes corresponded to that of global methylation averages. Consequently,

we did not detect a transcript class, whose transcripts were predominantly initiated from hypomethylated, reactivated promoters in Dnmt1$^{-/chip}$. Not even the promoters of the few truly differentially expressed transcripts [▷ Figure S9.1] exhibited a pronounced hypomethylation in Dnmt1$^{-/chip}$. Ultimately, convincing evidence for widespread reactivation of TINATs was lacking in our model and data.

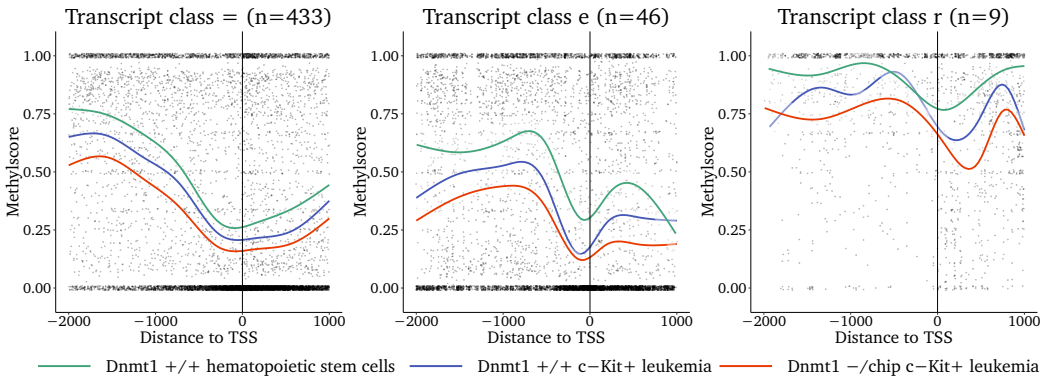


Figure S9.5: Measured (dots) and averaged smoothed (curves) promoter methylation of de novo assembled transcripts, which were differentially expressed between Dnmt1$^{-/chip}$ vs. Dnmt1$^{+/+}$. Colors represent the WGBS samples. Transcripts without WGBS data have been omitted.

9.4 Isolated transcriptional initiation events

The results presented in the previous sections of this chapter were based on the de novo transcriptome assembly generated from the merger of RNA-seq and CAGE-seq data integrated with WGBS data [↪ section 7.1, p.47]. However, it should not go unnoticed that the datasets were generated years apart and originated from different ex-vivo leukemia samples. Although the state of technology at the time did not permit to generate all data from the same replicates, reducing rather unrelated datasets to a common denominator bore some risk of loosing information, assuming at least some contribution of random demethylating events in shaping the transcriptome and methylome of Dnmt1$^{-/chip}$ cells. The paired-end CAGE-seq generated for the original TINAT publication [256] was more purposeful and straightforward in regard to the scientific question addressed.

Thus, technical limitations may have hindered us to capture the true extent TINAT-like transcriptional events in our mouse model, where DNA hypomethylation was archived by means of genic Dnmt1 reduction instead of inhibitor treatment. For example we failed to assemble a particular, truncated transcript of Pard6b, which Irina Savelyeva from our own laboratory had experimentally validated by 5′-RACE-PCR in several Dnmt1$^{-/chip}$ leukemia. Furthermore we also observed a strong CAGE-seq signal (and hypomethylation) right at the site of the cryptic promoter in the Dapk1 gene, which had been described in the original TINAT publication [256], but the RNA-seq data did not allow for a successful elongation into a de novo assembled transcript.

In a nutshell, our de novo assembled transcriptome of MLL-AF9 leukemia comprised mostly recurrent, faithfully reconstructed transcripts at the expense of most random, rare

RNAs originating from TINAT-like initiation events. Therefore, we once loosened the criteria and focused solely on the CAGE-seq data to elaborate on aberrant transcriptional initiation, although MLL-FP are known to rather affect the elongation of transcripts than their initiation [301].

We applied the software PARACLU [302] on the BBMAP [297] alignments [↪ section 9.1] to identify tag clusters. Such arrays of multiple CAGE signals in close spatial vicinity are representative of strong eukaryotic promoters, since most of them do not have a single transcription start site (TSS) but initiate RNA Polymerase activity in a narrow segment of the chromosome. The raw paraclu output was filtered[2] and intersected with the 652 852 mouse promoters released for *NCBI37/mm9* by the FANTOM 5 consortium [10, 303].

In total we could identify 140 267 tag clusters, of which 45 259 overlapped known promoters from the FANTOM 5 reference, while 95 008 were unique. We also filtered such tag clusters, which were rather enhancers than promoters according to our analysis. While the majority of annotated tag clusters were common to both genotypes, two-thirds of the unannotated sites were specific to either leukemia [▷ Table S9.2].

Specificity	Type	FANTOM 5 annotation	Counts
Combined	Enhancer candidate	annotated	1180
Common	TSS cluster	annotated	34 545
Dnmt1$^{-/chip}$	TSS cluster	annotated	6549
Dnmt1$^{+/+}$	TSS cluster	annotated	2985
Combined	Enhancer candidate	non-annotated	2050
Common	TSS cluster	non-annotated	32 320
Dnmt1$^{-/chip}$	TSS cluster	non-annotated	15 898
Dnmt1$^{+/+}$	TSS cluster	non-annotated	44 740

Table S9.2: Number of CAGE-seq tag clusters in each category. Since enhancers will be discussed later, the subcategories are not detailed here. Such clusters that have a match with the mouse promoter reference released by the Fantom 5 consortium [303] are considered to be annotated.

The most remarkable finding was the high number (44 740) of unannotated tag clusters in Dnmt1$^{+/+}$ MLL-AF9 leukemia. This result was very unexpected and would rather fit to the proposed changes in a hypomethylated setting. Thus, we double-checked the whole pipeline from barcode deconvolution to mapping and also performed a control alignment with BBMAP [297] of the single replicates ($n = 4$ per genotype) on the tag clusters. This alignment confirmed the expression exclusively in Dnmt1$^{+/+}$ and the near-equal contribution of all replicates. Consequently, if samples were confused, it did already occur in the lab during library preparation and likely affected more than two replicates. Unfortunately a retroactive bioinformatic validation was only feasible for the RNA-seq, but not the CAGE-seq samples [↪ subsection 7.2.2, p.50].

[2] Minimum of 15 reads total per cluster and an average of 0.5 reads per base to bias large sparse clusters.

At first, we explored the genomic localization of the tag clusters. For this reason we mapped the first principal component, which distinguishes *active/permissive* from *inactive/inert* chromatin compartments [↪ subsection 7.3.1, p.51] [38] , of Hi-C chromatin interaction data generated in the HPC-7 murine blood stem/progenitor cell model [↪ subsection 5.2.2, p.39] [246]. While basically all annotated tag clusters irrespective of their specificity were exclusively located in the open chromatin regions [▷ Figure S9.6, top row], we noticed an abnormal enrichment of Dnmt1$^{+/+}$-specific, unannotated TSS clusters in typically inert heterochromatic regions. Since the decompaction of chromatin is a prerequisite of active transcription, this either pointed towards an increased flexibility of the chromatin structure or a more readily initiated transcription [▷ Figure S9.6, bottom row].

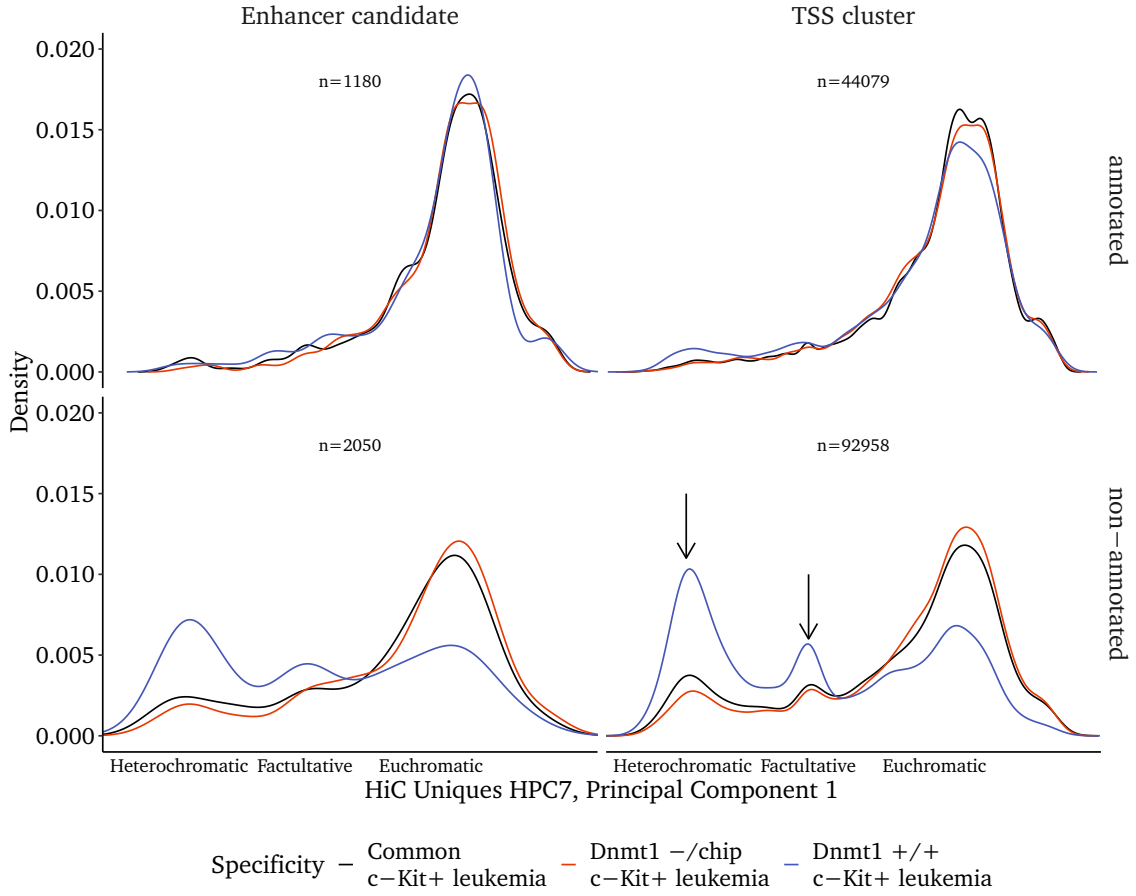


Figure S9.6: Genomic localization of transcriptional initiation. The tag clusters were assigned the respective first principal component of HPC-7 murine blood stem/progenitor cell Hi-C uniquely mapping interaction data. TSS were separated according to specific occurrence, overlap with Fantom 5 reference as well was classification as enhancer or promoter. Black arrows emphasize the unusual enrichment of robust or facultative heterochromatic localizations in the wild-type specific, unannotated clusters.

Next, we mapped the modeled methylation probability on the tag clusters to assess their most likely methylation state, since we lacked actual WGBS coverage for many of them. The shift in the samples' maxima reflected the average methylation levels observed by the 100 kb sliding window analysis [↪ section 2.0, p.17], but in general similar effects

were observed.

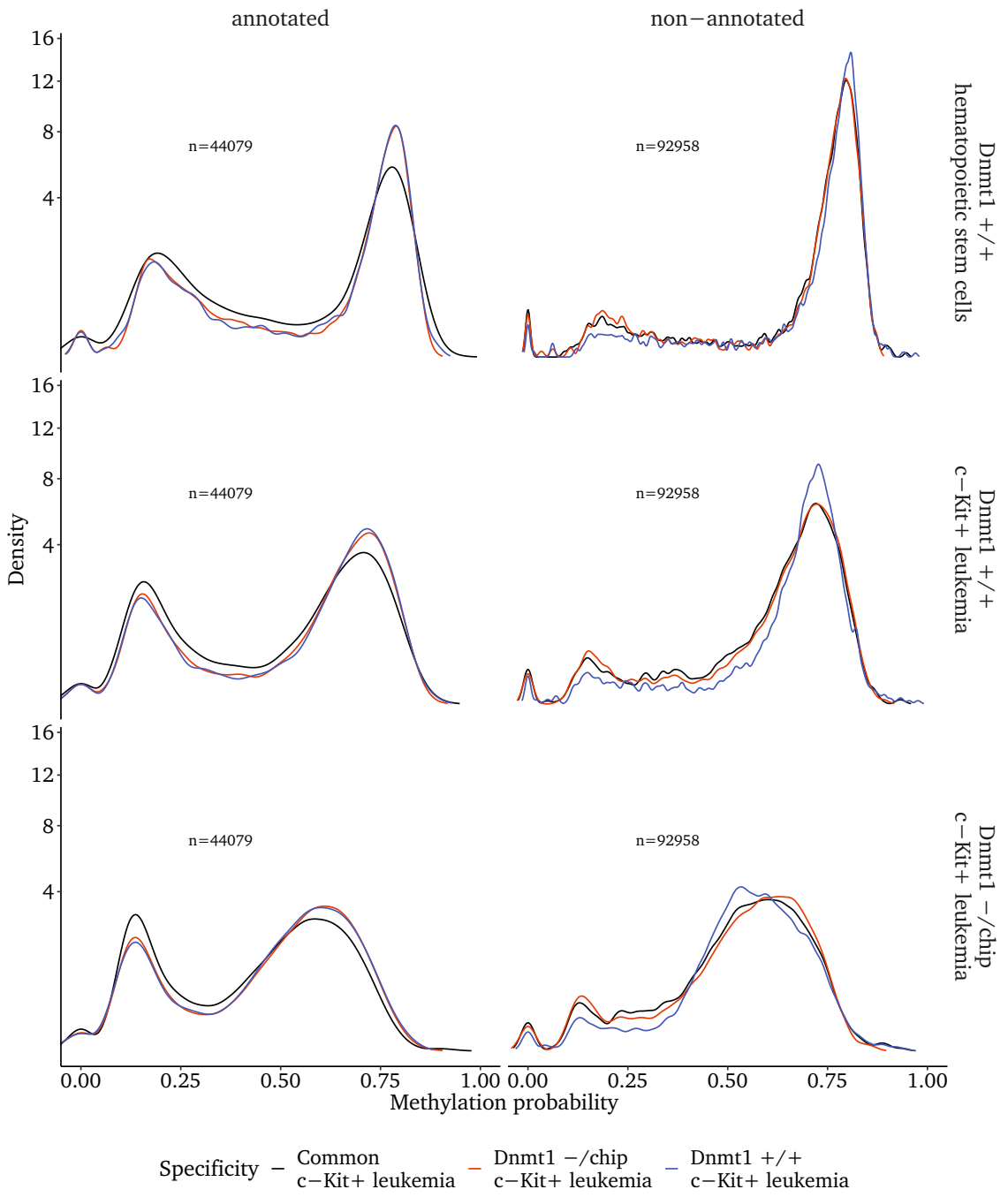


Figure S9.7: GAM-predicted methylation of CAGE-defined promoter TSS clusters (enhancers not shown in this plot). Kernel density estimates illustrate the frequency of the respective promoter methylation. Mind the non-linear y-axis required to represent the unequivocal methylation of non-annotated promoters in HSCs.

The annotated promoter TSS clusters split between methylated and unmethylated promoters, yet the common ones slightly favored a demethylated state [▷ Figure S9.7, left column]. Since the model was deliberately chosen to be very smooth and focused on the background methylation, values in the range of 0 to 0.25 were basically confined to CpG-Islands. Consequently, we could conjecture from this data that a low predicted

methylation was associated with CpG-Island promoters, which we corroborated in a separate analysis [▷ data not shown]. Conversely the non-annotated promoters were predominantly devoid of CGIs and thus they were typically methylated according to the model [▷ Figure S9.7, right column]. In addition, slight deviances were visible for the Dnmt1$^{+/+}$-specific non-annotated clusters: In Dnmt1$^{+/+}$, they exhibited the highest and in Dnmt1$^{-/chip}$ the lowest predicted methylation of all three specificity categories.

The latter was confirmed, when instead of the absolute methylation the relative differences between the two leukemia samples were determined: The Dnmt1$^{+/+}$-specific non-annotated TSS exhibited the most significant hypomethylation [▷ Figure S9.8, lower right panel], which was mirrored in the enhancers. Given their prevalent chromosomal localization in heterochromatic domains [▷ Figure S9.6, lower panels], the profound methylation loss was anticipated.

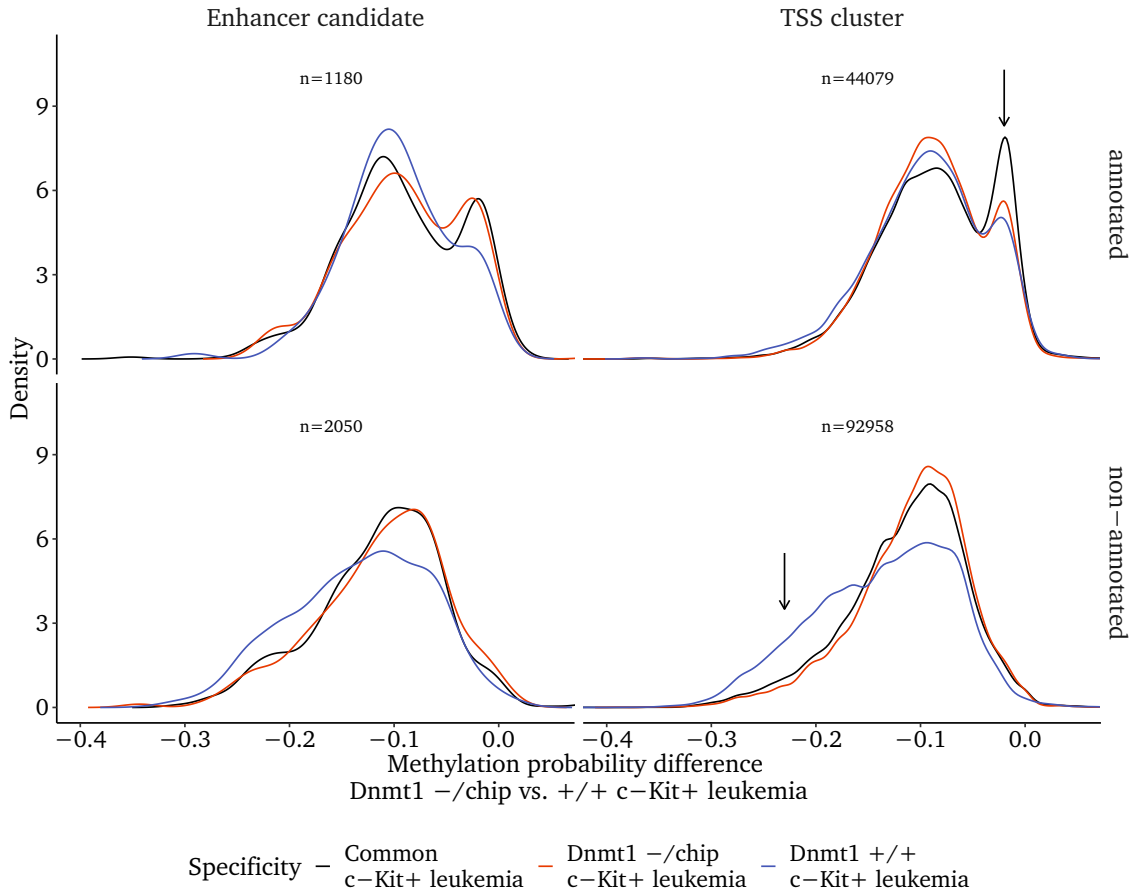


Figure S9.8: Relative GAM-predicted methylation differences for all possible combinations of annotation, cluster category and specificity. Black arrows point to the most relevant differences: The methylation loss of wt-specific non-annotated clusters and the elevated persistency at common, annotated promoters due to the high rate of CpG-Islands.

Supplementary Chapter 10

Enhancer calling and classification

10.1 CAGE-seq derived enhancers

Several methods exist, which can be used to identify putative enhancers in genome-wide datasets [reviewed in 133, 134]. Because we had already generated cap analysis of gene expression (CAGE-seq) [159, 160, 162] datasets from ex-vivo sorted the c-Kit$^+$ fractions of four independently established leukemia to characterize TINATs [↪ section 9.4, p.71], we ultimately settled for that approach.

CAGE-seq data versatilely allows for the parallel determination of transcript expression estimates, investigation of alternative promoter usage and the detection of sites with potential enhancer activity based on bidirectional eRNA transcription [↪ subsection 1.2.3, p.10]. A strategy to predict enhancers genome-wide from CAGE data alone has been devised, thoroughly validated and published in the course of the FANTOM projects [103]. We reproduced this approach as closely as possible, but replaced the aligner with BBMAP [297], since it allows for a magnitude more parameters to be set and thus to exert a fine-grained control over the process, which was required due to a higher sequencing noise in our data. Subsequently, we applied the software PARACLU [302] to identify tag clusters, which are arrays of multiple CAGE signals in close spatial vicinity. Such patterns emerge, because RNA polymerase activity is often initiated in narrow segments of the chromosome rather than at a specific single transcription start site (TSS). As recommended by the original authors, we excluded clusters found 500 bp or less away from known TSS or 100 bp from known exons. Source of the annotation used for the final reanalysis was release 84 of the NCBI REFERENCE SEQUENCE DATABASE (REFSEQ) published on September 11, 2017.

After reshaping the data to meet the input requirements, we executed the Perl script provided by the Andersson lab[1]. Since the programming interface of the BEDTOOLS suite [232], which is used by the script had changed in the meanwhile, slight modifications were required to run it successfully. Initially we obtained 8675 and 7955 bidirectionally transcribed clusters in Dnmt1$^{+/+}$ and Dnmt1$^{-/chip}$ respectively. We filtered such sites that were smaller than 80 bp or larger than 500 bp as well as such items, whose

[1] https://github.com/anderssonrobin/enhancers

cumulated expression did not exceed 0.5 TPM in total and 0.2 TPM in at least two replicates. These filters eliminated poorly supported locations with very weak signals, which were more frequent in Dnmt1$^{+/+}$ and thus corroborated the higher transcriptional noise in these samples [▷ Figure S10.1, black arrow in left panel]. Apart from the larger number of very weakly transcribed sites in Dnmt1$^{+/+}$, the distribution was similar in both genotypes [▷ Figure S10.1] and well in accordance with previously published data [103].

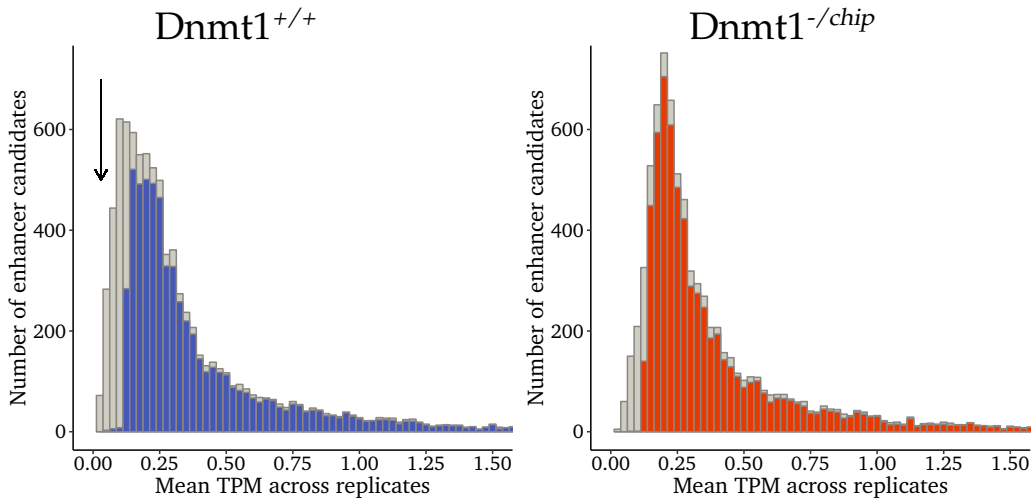


Figure S10.1: Histogram of bidirectionally transcribed enhancer candidates relative to the average normalized expression in tags per million (TPM). Shown in gray is the complete set, whereas the colored columns represent the final set that passed all filters. A vertical black arrow highlights the increased counts of weakly expressed, bidirectionally transcribed sites in wild-type prior to filtering.

However, it had already been shown in the supplement of the initial publication [103] that tightening the TPM cutoff beyond an optimum (which we determined to be 0.5 for our experiment) only eliminated many actual candidates without providing a significantly increased specificity. Therefore, we did not apply stricter filtering rules at this stage, deliberately accepting the likely presence of some false positives in the sets.

Next, we assessed, how heterogeneously the putative enhancer sites were transcribed in the biological replicates. For this analysis, we quantified the site's transcription in each replicate separately and ranked the measured values from large to small. Then, we calculated the cumulative sum of the two largest replicates per bidirectionally transcribed region. Assuming four exactly identical measured values, the sum of two would thus be 0.5, whereas the maximum value of 1 would indicate expression in just two out of four samples. Unsurprisingly the total expression of weakly transcribed sites was often dominated by one sample, although contribution by secondary samples was certainly present as they would otherwise have been excluded by the previous filtering steps.

Such sites may have either represented false positive calls or enhancers, which preferably resided in a silenced state in most cells [↪ subsection 1.2.2, p.9] and thus were of little relevance for leukemia. With increasing total expression, almost all bidirectionally transcribed sites were detectable in all replicates and the cumulative sum of two samples neared 0.5 [▷ Figure S10.2].

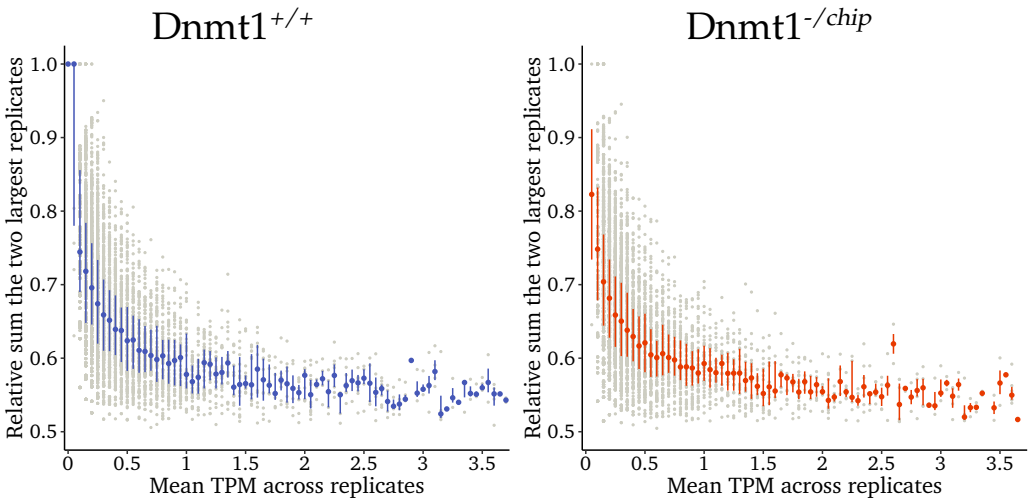


Figure S10.2: Heterogeneity of enhancer RNA transcription. The y-axis denotes the fraction of the total expression contributed by the two largest measurements. Individual enhancers (gray dots) are ordered from left to right by increasing mean expression and are binned by sequential steps of 0.05. For each bin the median of all enhancers is calculated and marked by a colored dot. The interquartile range (IQR) is shown as colored line.

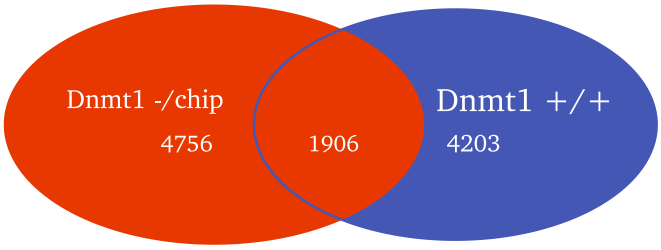


Figure S10.3: Venn diagram of the enhancer candidates, which passed the initial filtering step.

Ultimately, we retained 6386 and 6662 putative enhancers in Dnmt1$^{+/+}$ and Dnmt1$^{-/chip}$ respectively. Surprisingly, the majority of them (82.45 %) was specific for either of the genotypes [▷ Figure S10.3]. The large number of unconnected sites suggested a relevant share of false positive sites and called for extra caution in handling the data. Eventually, we gave consideration to means of increasing the reliability as well as validating candidate sites.

10.2 Enhancer intersection

Since we were working with leukemic cells, it seemed plausible that not all detected active enhancers would be involved in pathogenic progresses. Rather we expected to find hundreds of sites, which were inherited from the cell of origin and had no direct relevance for the leukemia. To segregate those two groups, we intersected the identified coordinates with the 48 415 enhancers contained in a comprehensive reference catalog of

murine hematopoietic enhancers [151].

The group of Ido Amit had developed a modified ChIP-seq protocol, which permitted the generation of sequencing libraries from just a few thousand cells. Subsequently, they had called enhancers by means of overlapping ChIP-peaks of H3K4me1 and H3K27ac. The extensive dataset, which comprised most major cell types spanning the whole hematopoietic hierarchy allowed them to identify clusters of congeneric enhancers. Overall, the group identified nine large clusters by *k*-means-clustering, which they designated with Roman numerals and which represent groups of functionally related enhancers [151]. While the enhancer coordinates were published in the supplement of the paper, the cluster assignments were kindly provided by Assaf Weiner and Ido Amit via direct correspondence. This allowed us to intersect the CAGE-defined candidate sites with the regular hematopoietic enhancers and also assign them to the previously defined H3K4me1-based clusters.

The cluster *Common* (I) comprised enhancers shared throughout hematopoiesis. Lineage-specific enhancers, which are already marked in HSCs and shared with hematopoietic lineage progenitors were grouped in the clusters II, III as well as IV. *Progenitors* (V) was shared mostly among progenitors and signal intensity decreased in the more mature cell types. The clusters VI-IX comprised mostly de novo enhancers that were specific to a particular lineage and were just weakly marked in HSCs [2].

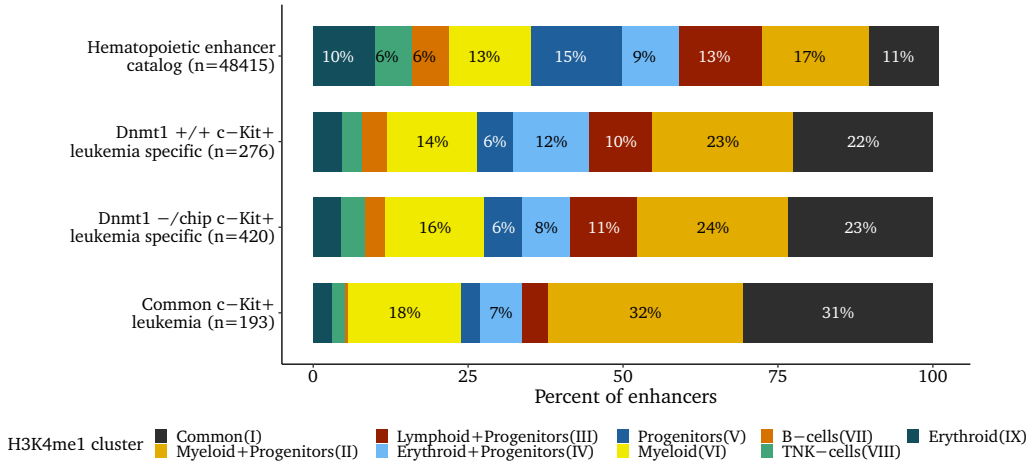


Figure S10.4: Bar graph showing the H3K4me1-cluster assignments of the overlapped hematopoietic enhancers. The top row serves as reference and represents the composition of the complete catalog. Shown below are three sets of regular hematopoietic enhancers, which are putatively also active in MLL-AF9 c-Kit$^+$ cells as well as their cluster allocation.

The result of the intersection gave us a first impression of the enhancer signature of c-Kit$^+$ cells. We could identify 889 CAGE-based enhancer candidates, which had been characterized before by the group of Ido Amit in healthy hematopoiesis. The percentage of normal enhancers in the signature (8.2 % of 10 865) was surprisingly low. Furthermore we noticed slightly disproportionate rates in the genotype-specific sets (Dnmt1$^{+/+}$: 6.6 %,

[2] see Figure S10.5,p.81 for a visualization of the clusters' H3K4me1 signatures

Dnmt1$^{-/chip}$: 8.8 %).

In terms of function, substantial fractions (37 % to 50 %) originated from the myeloid clusters *Myeloid + Progenitors* (II) and *Myeloid* (VI) [▷ Figure S 10.4], thus corroborated the known relatedness of the MLL-AF9 Lin$^-$ Sca-1$^+$ c-Kit$^+$ cells with granulocyte macrophage progenitors (GMPs), since C/EBPα-mediated differentiation is required for AML initiation [304]. Yet, a notable number of enhancers (up to 28 %) also originated from progenitor clusters of the other lineages (III,IV,V), which indicated that either lineage commitment had not been finalized or that the c-Kit$^+$-fraction consisted of a rather heterogenous mixture of various cellular stages.

10.3 Enhancer clustering

10.3.1 Major cluster assignment by k-means

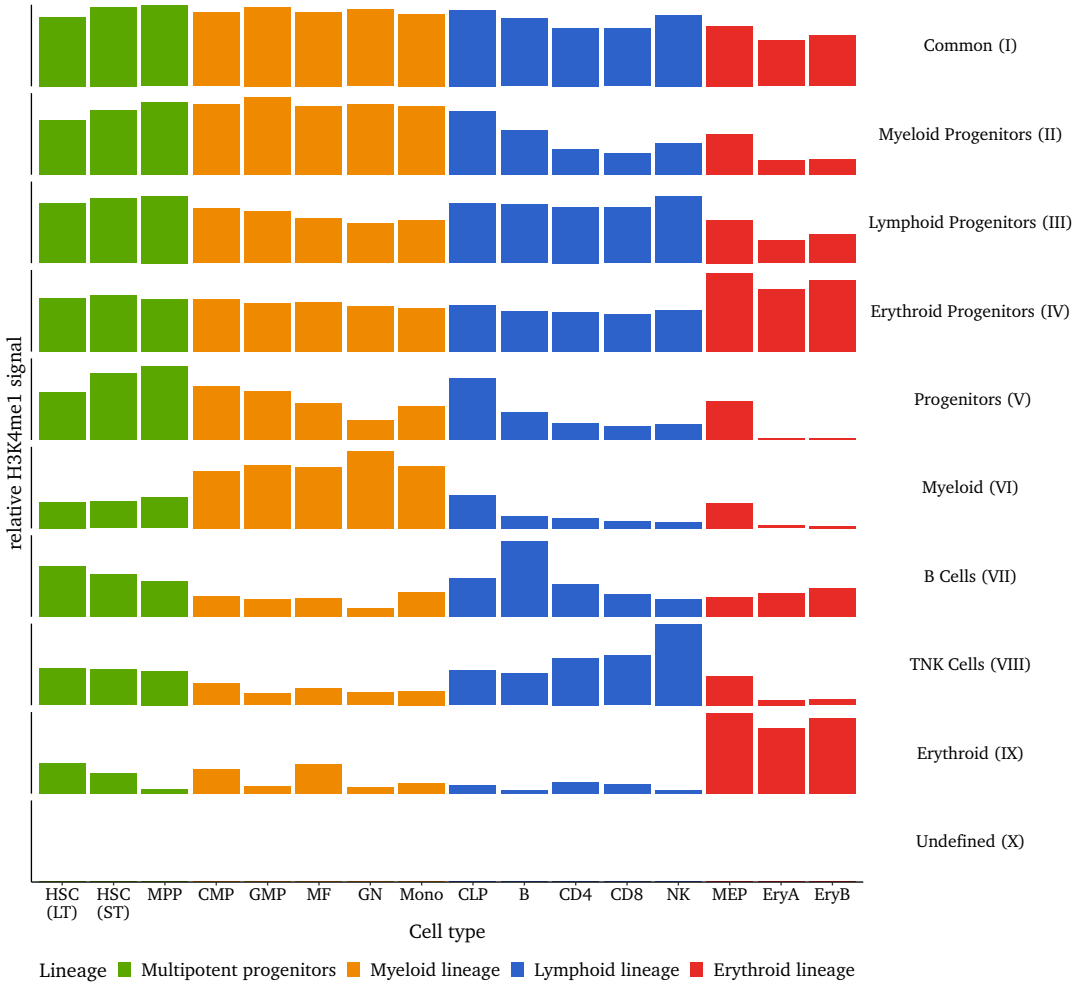


Figure S10.5: Bar graph showing the initial H3K4me1-signature associated with the cluster centers. For every cell type the mean value of H3K4me1-occupancy for all regular hematopoietic enhancers assigned to the respective cluster is shown. The *Undefined* (X) cluster is not part of the original publication.

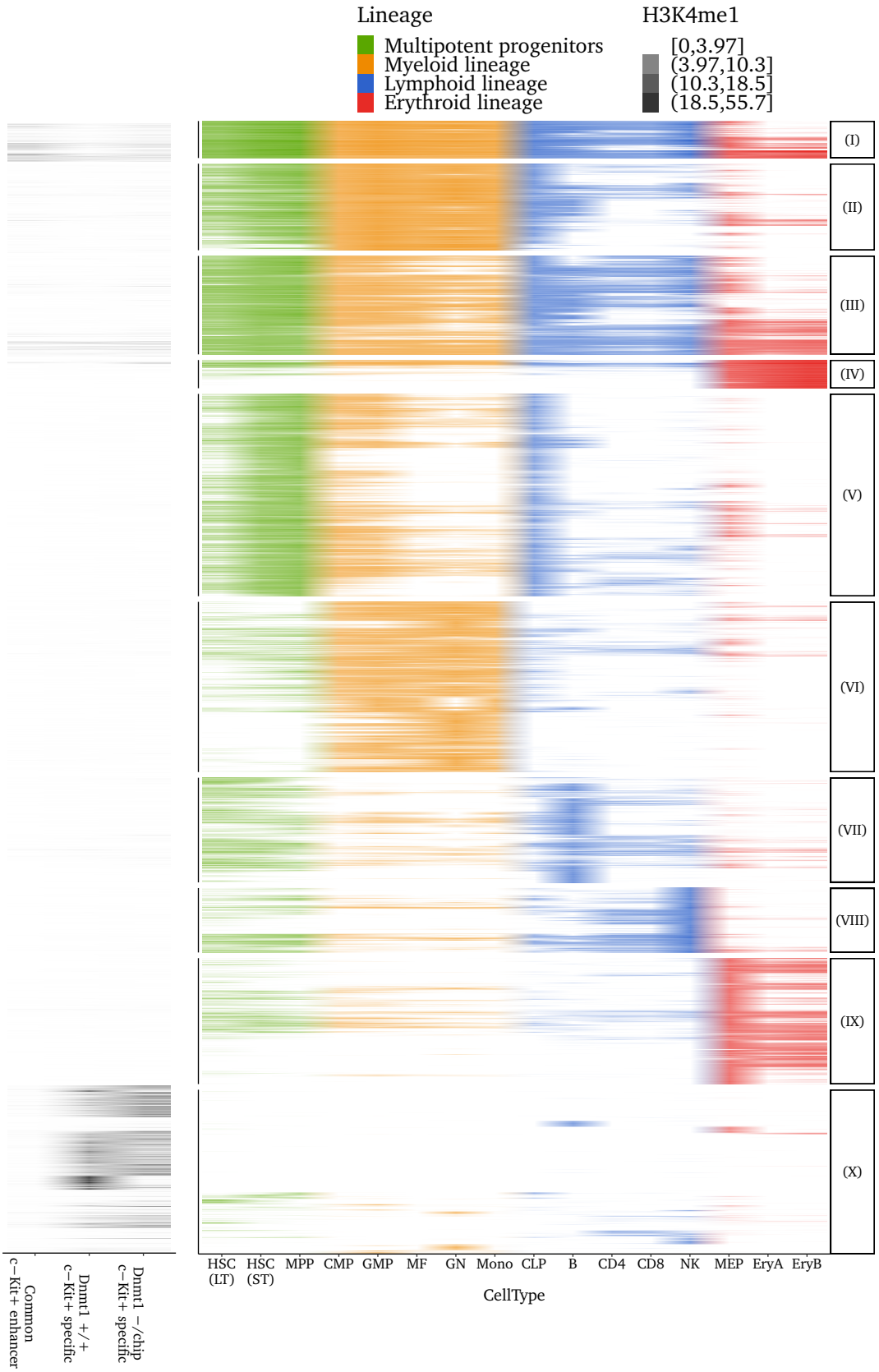
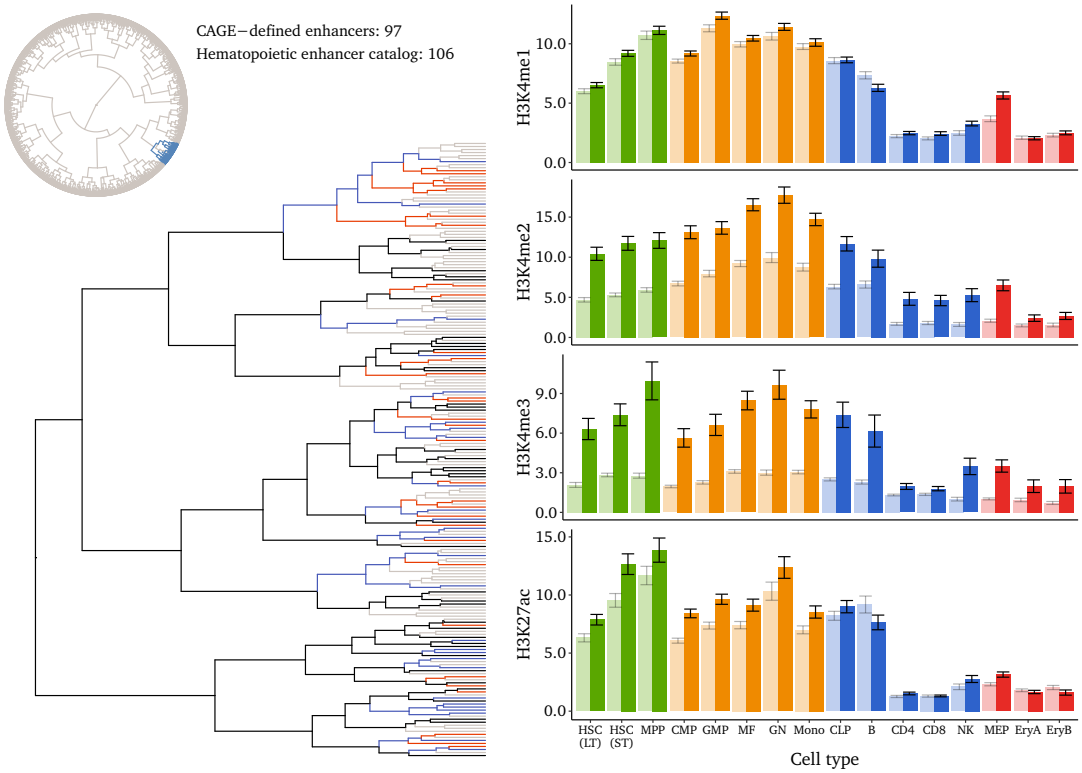
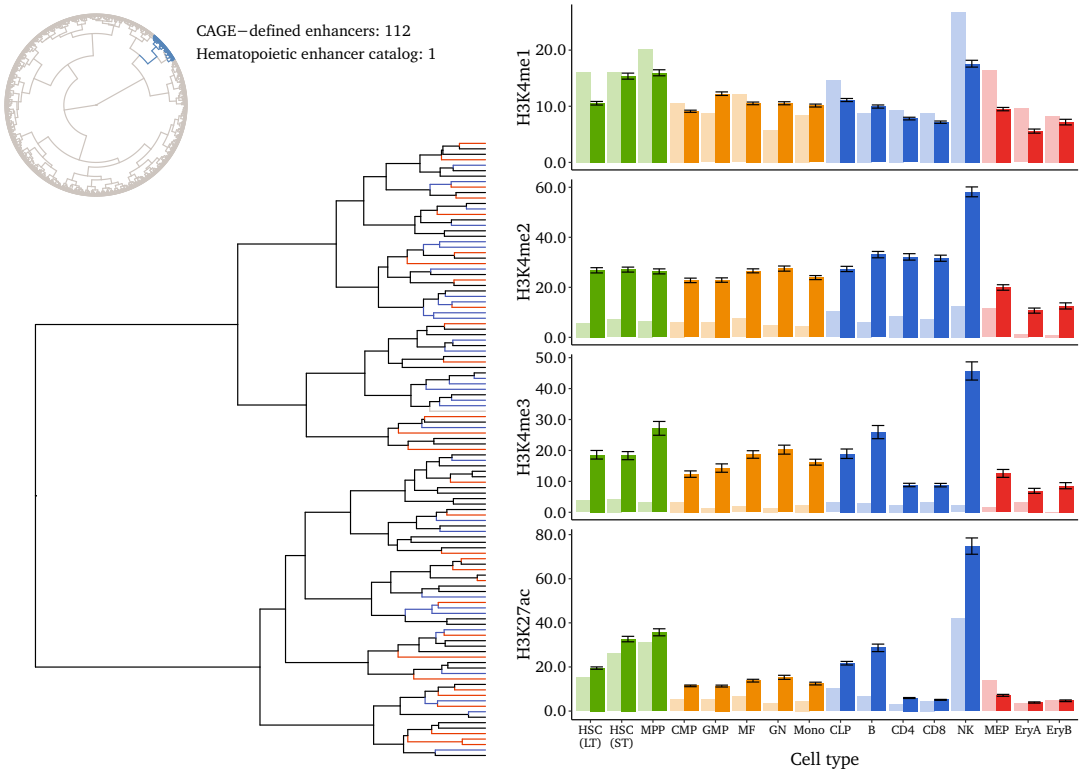


Figure S10.6: Heatmap showing the H3K4me1-occupancy in healthy hematopoietic cells at the sites of possible enhancers. All 10 865 candidates in MLL-AF9 c-Kit$^+$ leukemia as well as the 47 526 non-overlapping sites of the hematopoietic enhancer catalog are shown individually as horizontal lines. Roman numerals represent the clusters.

10.4 Additional plots for clades enriched for CAGE-enhancers



H3K27ac MyeloidProgenitors(II) clade 9



H3K27ac Common(I) clade 7

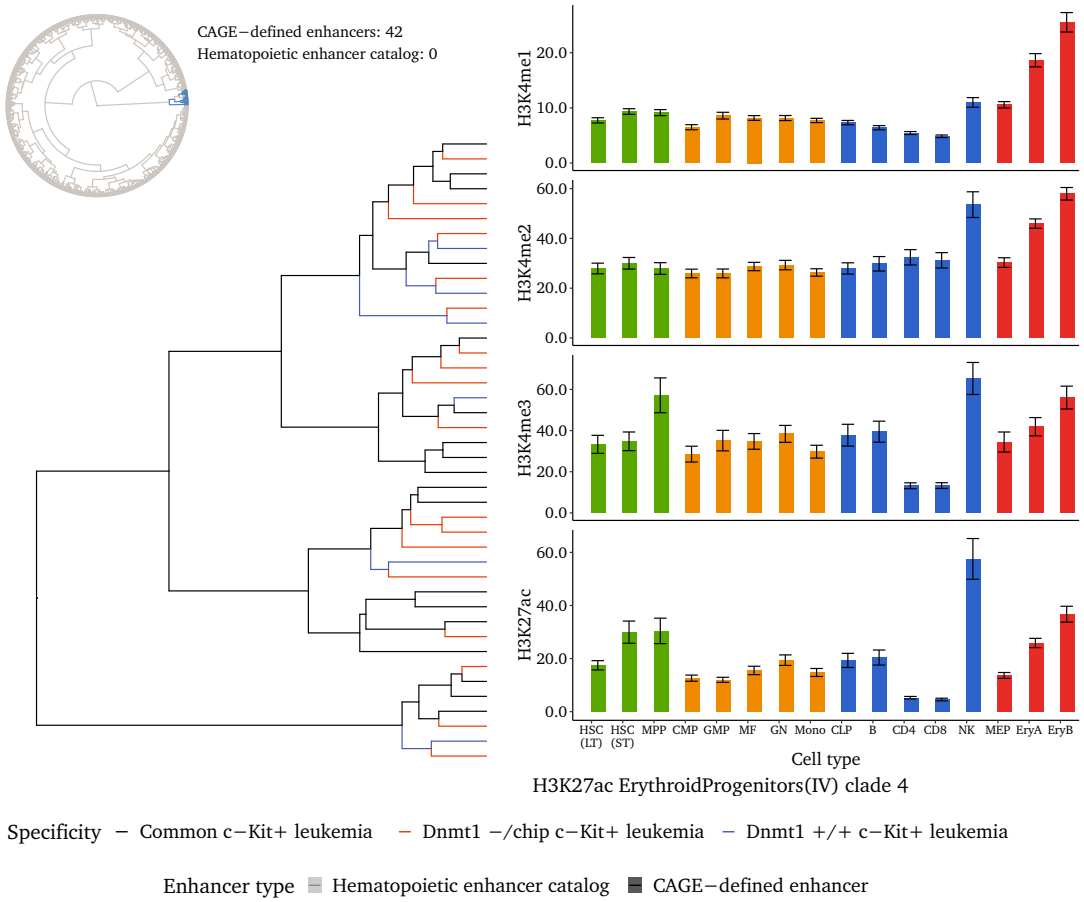


Figure S10.8: Details of three exemplary clades with enrichment for CAGE-defined enhancers (odds ratios 5.23, 77.37 and ∞) from the clusters *Common* (I), *Myeloid + Progenitors* (II) and *Erythroid + Progenitors* (IV). On the left the hierarchical clustering based on H3K27ac is shown as well as the genotype specificity: black marks common putative enhancers, blue Dnmt1 *+/+* and red Dnmt1 *-/chip* enhancers. On the right, note the strong H3K27ac signal at CAGE-defined enhancers (rich colored bars) in NK cells for the second and third clade. Although significantly enriched for CAGE-enhancers , clade *II.9* (odds ratio 5.23) lacks this prominent pattern.

10.5 ATAC-seq for enhancers in MLL-AF9 leukemia

ATAC-seq data generated by the group of Jennifer Trowbridge [305] also challenged the validity of most enhancer calls of the *Undefined* (X) cluster, as most of them were located in supposedly closed chromatin.

Apart from cluster X, the ATAC-seq however supported the notion that the clade enrichment was indicative of relevant enhancers. Typically, clades with a strong enrichment also featured elevated ATAC-seq signals. Furthermore the activity of such enhancers was ubiquitous in the sense that the openness of these chromatin regions did not depend on the cell of origin. Initially, the Trowbridge lab had transformed four distinct populations from the hematopoietic hierarchy by transduction of a MLL-AF9 fusion gene. The transcription profiles of the resulting leukemia were variable, but had no consistent rela-

tionship with the cell of origin, which was thus indistinguishable. However, the leukemia could be traced back to their cell of origin by means of open chromatin profiling [305]. Notwithstanding said cell type specificity reflected in the ATAC-seq data, the enhancers contained within the highly enriched clade were uniformly marked by strong signals and thus could be attributed to the core signature of MLL-AF9 leukemia.

However, the data also showed that the definition of such a core signature of MLL-AF9 leukemia would not be trivial and perhaps pointless. Even when just comparing the two biological replicates, the ATAC-peaks were quite variable. Therefore, a stringent intersection (reciprocal overlap of the peaks by at least 75 % of the length) eliminated many sites, including known promoters [▷ data not shown]. Cross-cell-type signatures were further narrowed down, such that the overall confirmation rate for the candidate enhancers by ATAC-seq was rather low [▷ Figure S10.9].

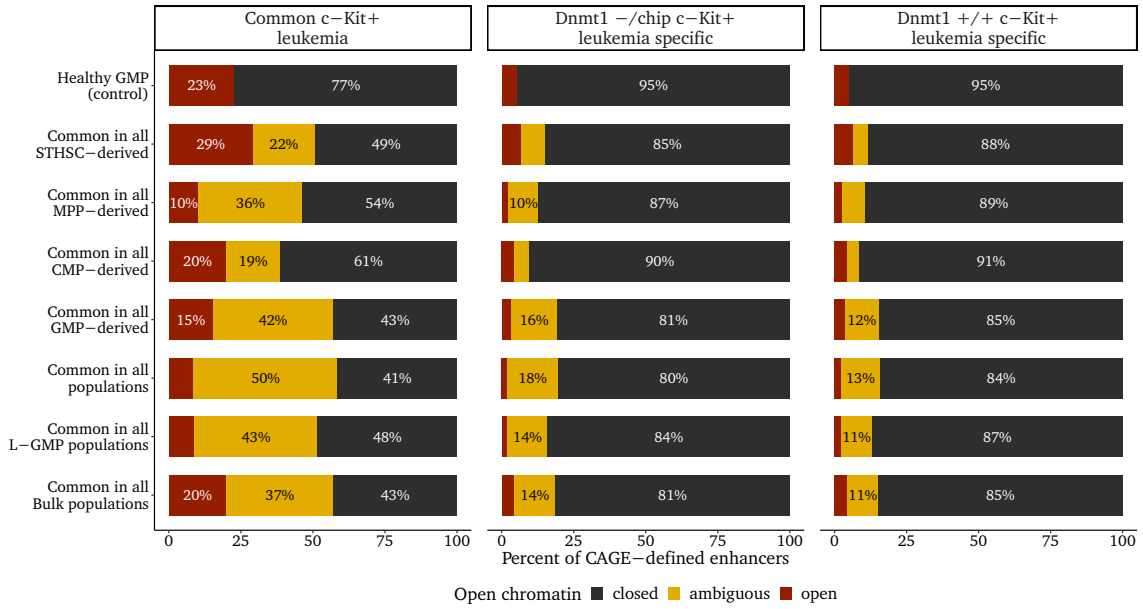


Figure S10.9: Fraction of CAGE-defined putative enhancers located in open chromatin according to ATAC-seq in the respective cell population. Peaks in replicates or mixed populations were required to overlap reciprocally by 75 % of the length to be considered commonly open. As no replicates were available for the healthy GMP control, the number of ambiguous elements could not be determined.

Among the candidates transcribed in Dnmt1$^{+/+}$ as well as Dnmt1$^{-/chip}$ 41 % of the elements seemed to be constitutively packed in heterochromatin in MLL-AF9 leukemia according to the ATAC-seq. Half of the candidates were open in at least some leukemic populations and just 9 % in all. Comparable numbers were obtained for the universal c-Kit^{high} LSC-enriched L-GMP (Lin$^-$ IL-7Rα^+ Sca-1$^-$ c-Kit$^+$ CD34$^+$ FCγR$^+$) signature. Judged on the basis of similarity to derivatives from a particular cell of origin, our Lin$^-$ Sca-1$^+$ c-Kit$^+$-derived c-Kit^{high} population predominantly resembled progeny from short-term HSCs, although they also bore some resemblance to the CMP-derived populations [▷ Figure S10.9, left column].

Confirmation rates were even lower for the vast majority of enhancer candidates, namely

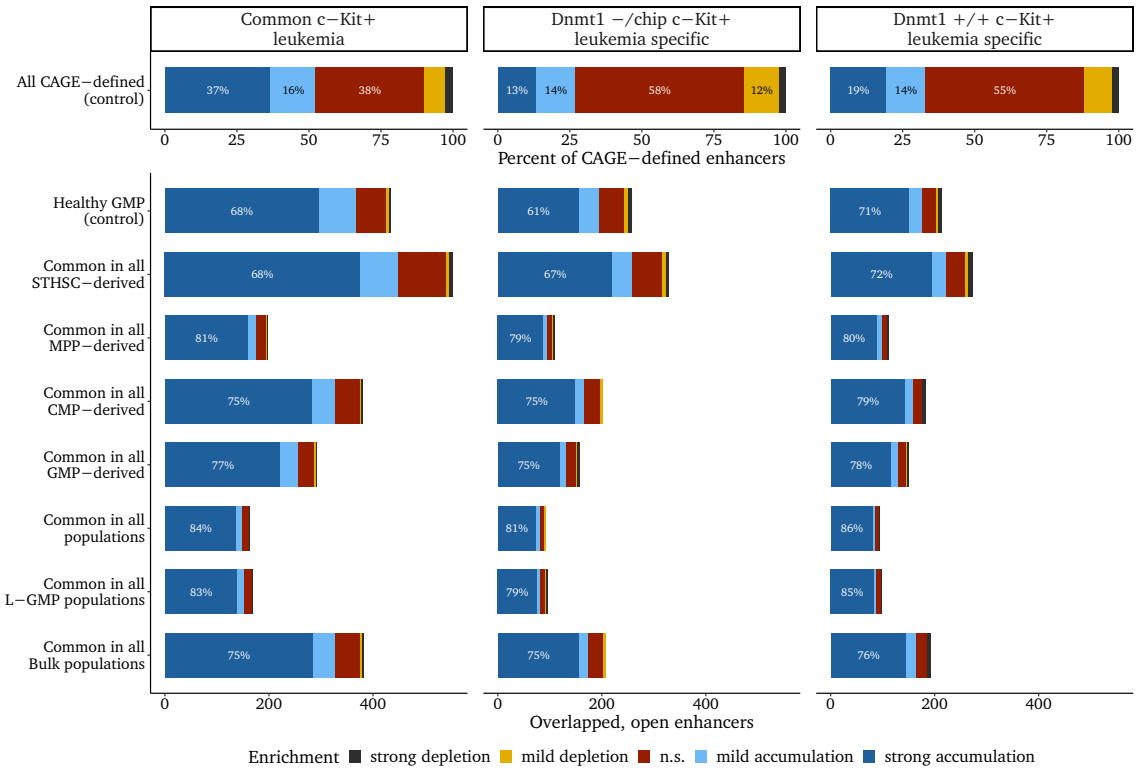


Figure S10.10: Absolute number of CAGE-defined candidate enhancers and their respective assignments to clades. Odds ratio (≤ 0.25; $]0.25, 0.75]$; $]0.75, 1.25]$; $]1.25, 4]$; >4) as well as test significance (FDR < 0.01 determined the five categories ranging from depletion to enrichment. Only those enhancers are shown, which are confirmed by ATAC-seq open chromatin profiling as constitutively open. On the very top, the shares to the respective clade categories in the full set are depicted as control - due to disproportional numbers.

82.45 % [▷ Figure S 10.3, p.79], which were specific for either genotype. Typically, just 15 % or less were assigned to open chromatin, a minority of which could be reckoned constitutively open [▷ Figure S10.9, center and right column]. Surprisingly, the higher rate of heterochromatic transcription initiation events among the Dnmt1$^{+/+}$-specific enhancers [▷ Figure S 9.6, p.73] was hardly reflected in the ATAC-seq confirmation rates - only 2 % to 4 % more open chromatin was detected for the Dnmt1$^{-/chip}$-specific over the Dnmt1$^{+/+}$-specific enhancers.

Regardless of the set and the level of confirmation, the majority of confirmed putative CAGE-defined enhancers were assigned to clades exhibiting strong accumulation. Apart of the healthy GMP control, where it was lower, between 67 % to 86 % of the confirmed enhancers were constituted by strongly enriched clades [▷ Figure S10.10].It is important to emphasize that this was not the very same group of enhancers, which affected the results over and over again in various contexts. Instead, we observed a quite heterogeneous mixture of ubiquitous as well as more specific enhancers, yet commonly assigned to accumulated clades, which were apparently involved in the leukemic transformation [▷ Figure S 10.11]. Therefore, we hypothesized that we should focus on clades with strong accumulation to investigate possible mechanisms governing enhancer recruitment

in MLL-AF9 leukemia .

In parallel, we also pursued an alternative approach to the identification of commonalities between those enhancers: We launched a search for distinct cis-regulatory elements, which presumably support the expression of genes that have a known role in MLL-AF9 leukemogenesis by harnessing chromatin interaction data.

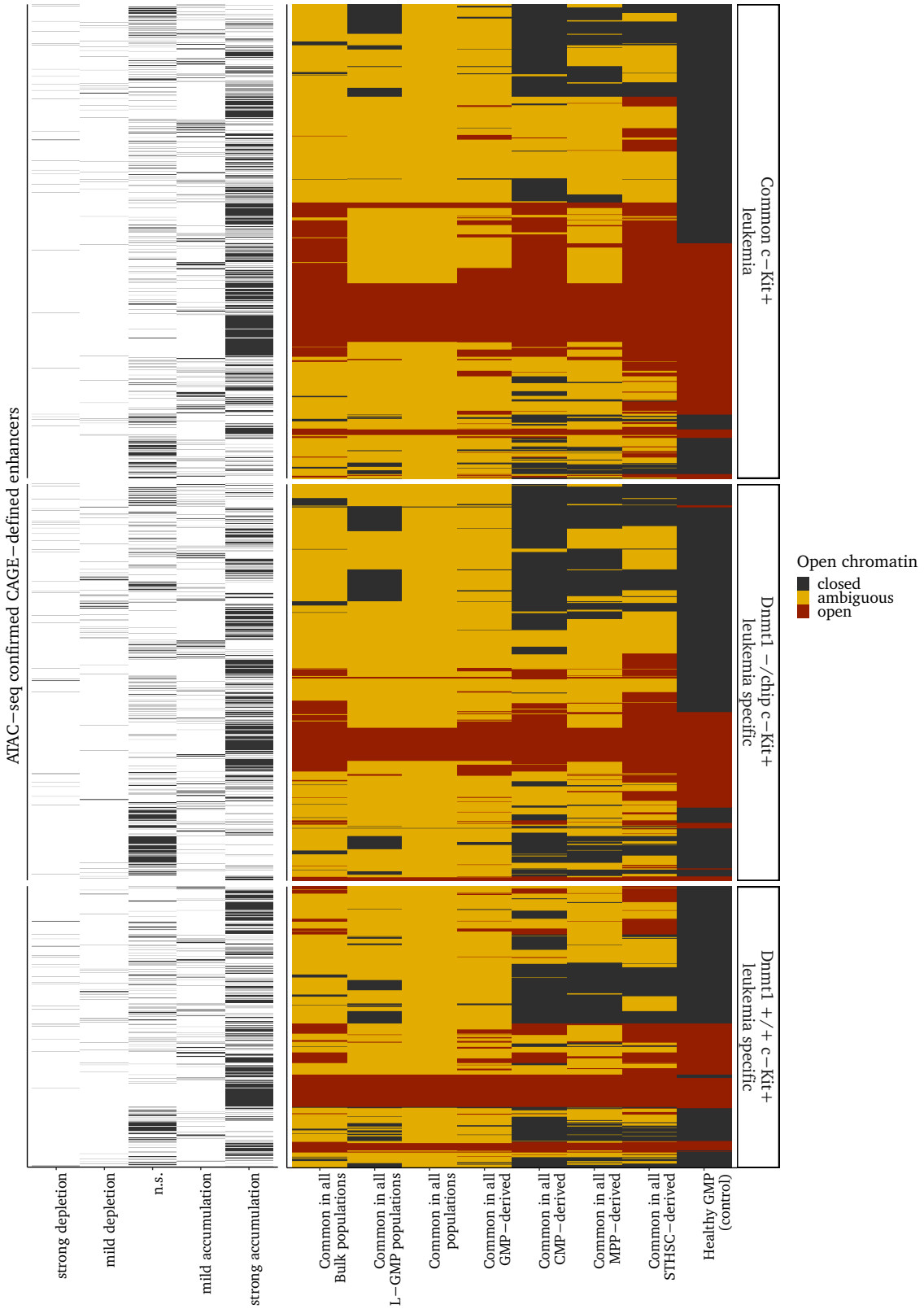


Figure S10.11: Heatmap showing the clade category and chromatin status of all ATAC-seq confirmed enhancers. Displayed are only such CAGE-defined enhancers, which were marked by an ATAC-seq peak in any sample ($n = 2716$). In the left panel, a black bar indicates the clade category, which is determined by odds ratio and false discovery rate as described before (Figure S10.10). A heatmap to the right details the presence of ATAC-seq peaks at the enhancer sites in the respective set.

Supplementary Chapter 11

Enhancer motifs and regulation

11.1 Methylation of enhancers and their motifs

DNA methylation is a crucial regulatory layer for normal and malignant hematopoiesis [reviewed in 306, 307] and a growing body of papers stress the importance of influential methylation changes at cis-regulatory elements in health and disease [63, 233, 308–311].

Identification of a potential CXXC zinc finger motif within the strongly accumulating enhancers suggested an investigation of the methylation status, since CXXC binds exclusively to unmethylated CpG-dinucleotides [312]. Also many other transcription factors are known to bind in a methylation sensitive manner [313, 314]. Decisive regulatory methylgroups do not necessarily have to be located directly at the binding site of the transcription factor: A particularly interesting paper had shown how methylation at distant sites facilitates the efficiency of Egr1 target search process [315].

Therefore, we hoped that decisive methylation changes in those regulatory regions might be the long sought answer to explain the Dnmt1$^{-/chip}$ phenotype, in particular its self-renewal bias observed in leukemic stem cells (LSC) [316].

11.1.1 Methylation mapping at isolated motifs

None the less, we hoped that we could manage to identify individual transcription factor binding motifs that might be disturbed by abnormal methylation. However, investigation of the methylation status was challenging, because confounding motifs had to be considered.

Contrary to promoters, motifs are small and rarely occur isolated in the genome, but typically appear in dense clusters. Therefore, the methylation status of a CpG can theoretically be shaped by the motif under investigation as well as neighboring ones. To account for this properly, we first searched for frequently co-occurring motifs and then tried to model their reciprocal influence.

For this task we employed the ARULES R package [317], which provides a computational environment for association rule deduction and frequent item set identification. Such

analyses are commonly used for finding regularities in the shopping behavior of customers. It aims to find sets of products, which are frequently bought together to stimulate purchases by improved arrangement in shelves, cross-selling or product bundling. However, these algorithms can also serve the purpose of identifying linked motifs by treating a motif instance as item and a cis-regulatory element as transaction set [318,319]. We tested several approaches and ultimately settled for the APRIORI algorithm [320] to derive frequent motif patterns.

Remarkably, the motifs ***DeNovo.***YAAAAAAAAV, ***DeNovo.***SSCGCGGCCTSS and ***DeNovo.***CCGGRCGGCG were strongly associated with themselves, as such they frequently appeared two to three times per enhancer. Minor associations were found with ***Nkx3.1.Homeobox.****GSE28264* and ***RUNX.AML.****Runt.CD4*.

A distinct, second set involved ***DeNovo.***CACTTCCTGT and ***PU.1.ThioMac.****PU.1.ChIP.Seq.Homer*. This was to be expected, because one had to assume on the basis of the high sequence similarity that it was de facto the same motif. Weaker associations of ***DeNovo.***CACTTCCTGT existed with ***Ets1.like.CD4.****PolII.ChIP.Seq.Homer* and ***MYB.HTH.****GSE22095*. The latter was also involved in a third set together with ***SCL.bHLH.****GSE13511* and ***DeNovo.***TGACGTCACT. Occasionally, also the motif ***Nanog.Homeobox.****GSE11724* appeared in this set.

The reason, why suddenly novel hits appeared in this analysis, which were not found previously, was the inclusion of a relatively large region of 4 kb as a consequence of the broad demethylation observed around the enhancer regions of strongly accumulated clades. Therefore, promoter motifs of nearby transcription start sites (TSS) may have been incorporated into the mined motif sets.

Subsequently, we tried to model the methylation change of the motifs with various generalized additive models (GAM), which incorporated the presence of other motifs from the frequent item sets as well as previously identified predictors such as DNA sequence or chromatin structure [↪ section 5.2, p.35]. Yet, despite comprehensive modifications at the models' terms, we did not succeed to come up with satisfactory predictions [▷ data not shown].

To some degree, this may have been due to the extremely bad coverage, which we recorded at many motif regions. For only 14 (2.15 %) respectively 15 (2.3 %) out of 651 instances of ***Ets1.like.CD4.****PolII.ChIP.Seq.Homer* in CAGE-defined putative enhancers, we had methylation information available in Dnmt1$^{+/+}$ and Dnmt1$^{-/chip}$. Even in cases where we achieved exceptionally high coverage, it was still mostly in the single-digit range. We obtained methylation information for 80/1313 ($\approx$ 6 %) instances of ***CEBP.CEBPb.****ChIP.Seq.Homer* and for $\approx$ 1750/8076 (21.67 %) of ***DeNovo.***SSCGCGGCCTSS. This low rate was particularly problematic for the proper consideration of co-occurring motifs, because there was virtually no set in which methylation calls for all involved motifs were present. Thus, we were not able to elucidate the reciprocal influence of frequently associated motifs on the respective methylation.

Ultimately, we resorted to the fits of the Kumaraswamy distribution [229, 230], a beta-

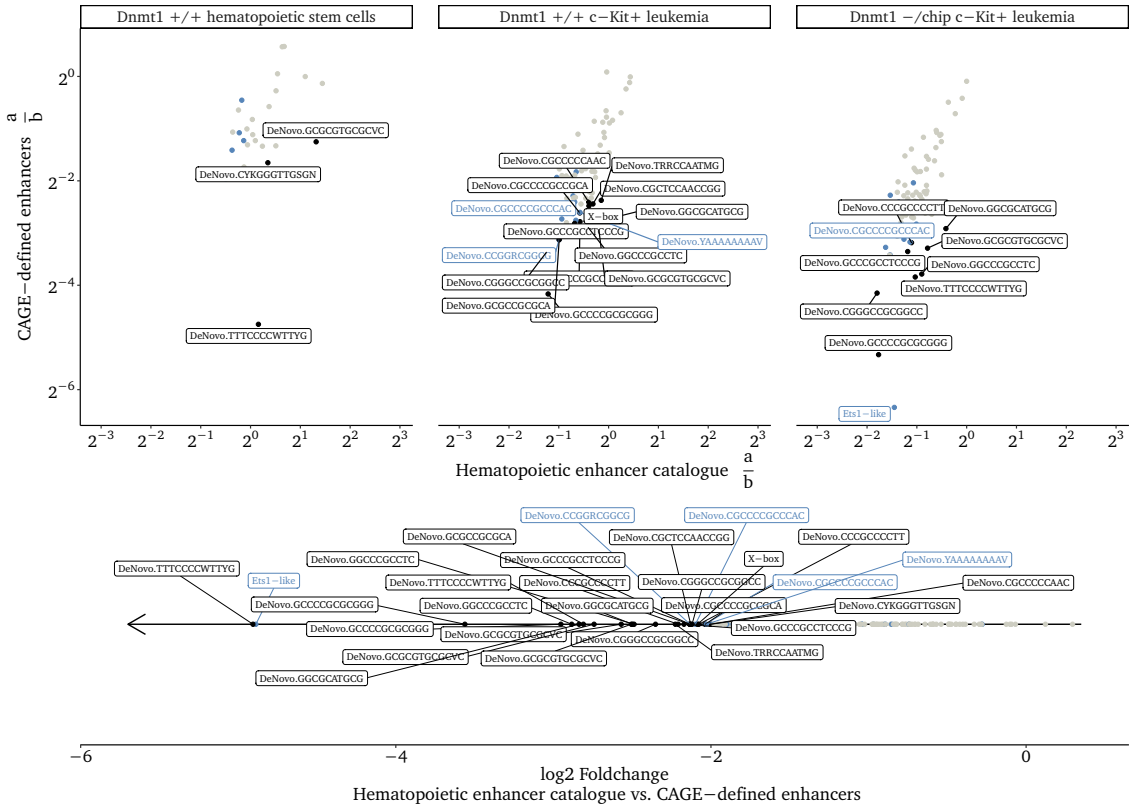


Figure S11.1: Two alternative representations of the screening results for methylation-sensitive motifs. Each motif is represented as a dot, but just the most relevant ones, which differed strongly between CAGE-defined enhancers and control, are fully labeled. The rest is shown as gray dots only. Blue color represents those motifs, which account for the top ten in strongly accumulated clades. In the top panel, the $\frac{a}{b}$ ratios of the motif-wise Kumaraswamy fits are plotted on the axes - either that of the CAGE-defined putative enhancers or that of the control. In this representation, the three WGBS datasets are shown separately, whereas the panel below does not discriminate the samples. Here, motifs are ranked according to the foldchange calculated as shown in Equation 11.1. Small $\log_2 FC$ indicate motifs, which are demethylated in the CAGE-defined putative enhancers but methylated in the control sites.

type distribution with more convenient tractability [↪ subsection 3.3.2, p.22], to assess methylation changes at the motifs and neglected possible combinatorial interactions. The shape of its probability density function (PDF) [▷ Equation 3.4, p.23] is determined by two parameters designated $a > 0$ and $b > 0$. The ratio of the two reflects the distribution of methylscores in the interval $[0, 1]$. If $a > b$, then the motif tends to be methylated, whereas $a < b$ will suggest a preference for unmethylated instances.

For all motifs, we thus calculated the ratio of $\frac{a}{b}$ for the motifs within the CAGE-defined putative enhancers as well as for inactive control sites from the hematopoietic enhancer catalog [151] [▷ Figure S11.1, top row]. Virtually all motifs exhibited a lower $\frac{a}{b}$ ratio for the CAGE-defined candidate sites and thus were demethylated in comparison to the control group. To represent this property in one dimension, we calculated the foldchange of the two ratios as shown in Equation 11.1. Motifs with a $\log_2 FC < -2$ were considered to be of particular interest due to the significant change in methylation [▷ Figure S11.1,

bottom row] and subsequently investigated in greater detail.

$$\log_2 FC \;=\; \log_2\left(\frac{a_{\text{CAGE}}}{b_{\text{CAGE}}}\right) - \log_2\left(\frac{a_{\text{control}}}{b_{\text{control}}}\right) \;=\; \log_2\left(\frac{a_{\text{CAGE}} \cdot b_{\text{control}}}{a_{\text{control}} \cdot b_{\text{CAGE}}}\right) \tag{11.1}$$

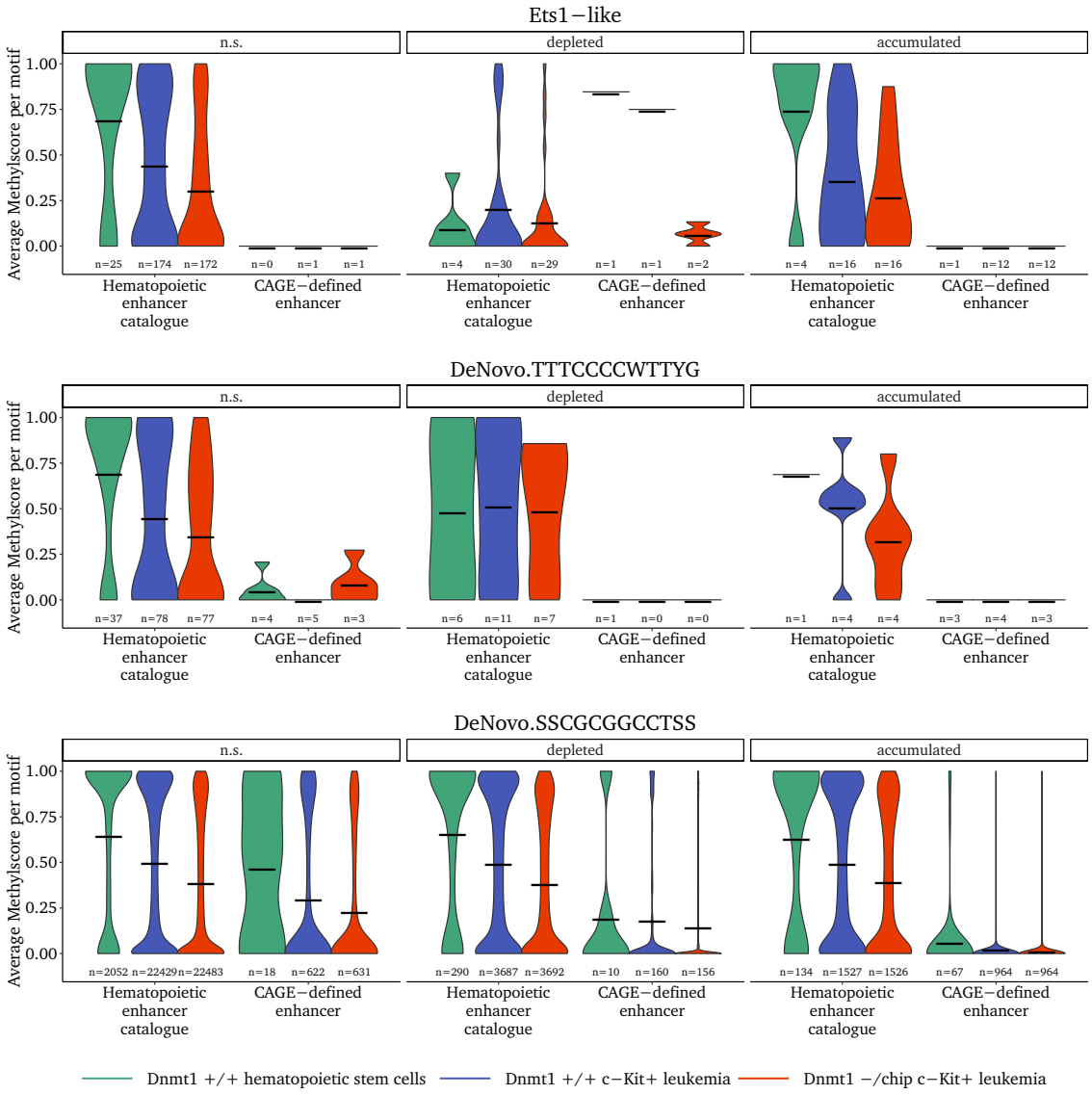


Figure S11.2: Detailed representation of three selected motifs and their methylation dynamics. For this plot, motif instances had been split among the enhancer groups and the WGBS meta-samples were mapped. The average methylation was calculated per motif instance and all covered sites (see counts below) were included in the violin plots. Methylscore distributions are depicted as vertical density plots and the methylation mean of the respective motifs as horizontal black bar.

The clearest demethylation was found for the motifs ***Ets1.like.CD4.***$_{\textit{PolII.ChIP.Seq.Homer}}$ as well as ***DeNovo.***$_{\text{TTTCCCCWTTYG}}$ [▷ Figure S 11.2, top and middle row] . However, both motifs were sparsely covered in the WGBS data, plus ***DeNovo.***$_{\text{TTTCCCCWTTYG}}$ was relatively rare ($n_{\text{CAGE-defined}}$ =227, n_{control} = 5463). Nevertheless, the counts were commensurate, since ***Ets1.like.CD4.***$_{\textit{PolII.ChIP.Seq.Homer}}$ was about three times more frequent in both categories

($n_{\text{CAGE-defined}}$ =651, n_{control} = 16663). Assuming that the few covered instances of the two motifs are representative, we observed an almost complete demethylation in CAGE-defined enhancers, but an ambiguous methylation in the controls. Here, an analysis of co-occurring motifs would have been of great interest, but was not feasible due to the low coverage. Methylation patterns of ***Ets1.like.CD4.***$_{PolII.ChIP.Seq.Homer}$ were very similar to ***PU.1.ThioMac.***$_{PU.1.ChIP.Seq.Homer}$, corroborating a potential functional equivalence in leukemia [▷ data not shown].

DeNovo.$_{\text{SSCGCGGCCTSS}}$, the most frequent motif in CAGE-defined putative enhancers assigned to strongly accumulated clades, was covered in 21.67 % of said enhancers in leukemia. At large, all instances within active sites were unmethylated in MLL-AF9 leukemia, but still partially methylated in hematopoietic stem cells (HSCs) [▷ Figure S 11.2, bottom row]. Remarkably, the degree of methylation HSCs, but not in leukemia, varied depending on the clades. The motif was mostly demethylated if found in the accumulated clade enhancers and ambiguously methylated elsewhere in CAGE-defined enhancers. In the control set derived from the hematopoietic enhancer catalog, the motif was ambiguously methylated, too. Generally, it exhibited the highest average methylation in HSCs and the lowest in Dnmt1$^{-/chip}$ leukemia. Clade enrichment did not matter for the methylation of control enhancers.

This pattern suggested an active regulation of the motif's methylation in the hematopoietic system. Therefore, it was intriguing to speculate that the variable methylation might alter the binding of a protein and we aimed to identify said protein.

Supplementary Chapter 12

Enhancer target genes

Contents

Elucidation of the mechanistic commonalities of the recruited enhancers was only one of two strategies, which we followed up in parallel. The basis of a second approach was the identification of target genes, which would permit to single out highly relevant enhancers by their target genes.

In scientific literature, many genes linked to the development and expansion of AML have been described. Examples include the homeobox cluster (HOX) [321,322], epithelial-mesynchymal-transition-related genes [282] or selected E3 ubiquitin-protein ligases like Xiap [323, 324]. We inspected such known proleukemic genes in our MLL-AF9 mouse model and tried to identify the relevant cis-regulatory elements sustaining their expression. Subsequently, we selected a small set for validation by CRISPR-Cas9 mediated silencing.

12.1 Assignment of enhancer target promoters

Establishment of the enhancer-gene connections was the most challenging part of task, since the closest gene respectively promoter is not necessarily the regulated one [325]. Furthermore, one transcription start site (TSS) is typically targeted by several enhancers, while one cis-regulatory element may also be involved in the regulation of different genes [17–19].

For the establishment of reliable enhancer-promoter interactions, we counted on published experimental chromatin interaction data rather than just assigning the closest transcription start site. As mentioned previously, the group of Berthold Göttgens gener-

ated Hi-C chromatin interaction data of the HPC-7 murine blood stem/progenitor cell model [246]. Part of this dataset, which we deemed to be a suitable proxy for the chromatin conformation of our MLL-AF9 c-Kit$^+$ leukemia model [↪ subsection 5.2.2, p.39], was a promoter-capture Hi-C experiment. A Capture-Hi-C protocol differs from a regular proceeding by an additional probe-based enrichment step to increase the coverage of predefined regions. The Göttgens group used this technique to specifically assay interactions with promoters of known transcripts.

We utilized this Capture-Hi-C dataset to map potential interaction partners of the CAGE-seq defined putative enhancers. In the course of the experiment, the Göttgens laboratory had fractionated the genome with the restriction enzyme *HindIII*. Thus, we in-silico digested the *NCBI37/mm9* reference genome accordingly to generate a fragment library to map the interactions to.

Even under optimal conditions, the resulting fragment sizes represented the technical limit in terms of resolution, which was achievable by this type of experiment. If a fragment contained several promoters or cis-regulatory elements, it was impossible to assign the interaction to a particular genetic element. Therefore, we counted the reads connecting two fragments and divided the sum by the number of contained elements. This approach allowed to assign a score to every possible connection, which represented the respective confidence in being a true interaction.

Additionally, we incorporated the FOCS [326] enhancer reference of validated enhancer-promoter pairs by increasing the score by 5 in case of a referenced pair. Importantly, we just altered the confidence score of an interaction, but did not introduce pairings from other sources, which were not supported by the HPC-7 Hi-C data. In total, we could derive 11 534 potential pairs, comprising 3103 putative enhancers and 4317 genes (6728 transcripts) from release 84 of the NCBI Reference Sequence Database (RefSeq) published on September 11, 2017.

Subsequently, we integrated the enhancer-promoter pairs with the RNA-seq expression data, which comprised 9505 expressed[1] transcripts [↪ section 7.3, p.51].

For the majority of transcripts, which were expressed according to RNA-seq, we could not assign a possible CAGE-defined enhancer from our set (n = 6856, 72.13 %). This, however, was not synonymous with a lack of any enhancers: First and foremost, the Hi-C method only captures interactions of different genomic fragments and the analysis strategy, which we applied, fosters the identification of far-cis interactions at the expense of near-cis interactions. Therefore, short-range enhancer-promoter-contacts, which are generally more frequent [165, 327], were underrepresented relative to long-range connections. On top of that, only transcribed enhancers are captured by CAGE-seq and enhancer-RNAs (eRNAs) are quickly degraded compared to mRNA so we possibly missed out on transient enhancer activity.

Conversely, roughly 40 % of the transcripts, to which an enhancer could be assigned, were

[1] at least 2 CPM in two or more single samples

expressed according to RNA-seq ($n = 2849$, 42.34 %) . While this might seem like an unacceptably high rate of false positives, it should be viewed in the light of enhancer activity preceding transcription [110] and multiple, alternatively spliced transcripts, which originate from the same or nearby promoters. Therefore, the proportion of expressed genes among those for which an enhancer interaction was predicted was higher ($n = 2750$, 63.7 %) and would have surpassed 80 %, if low-confidence interactions had been eliminated beforehand. Confidence scores ranged from 0.03 to 123.70, but we ultimately refrained from choosing a hard cutoff to eliminate false positive connections because lacking experimental validation any threshold would have been essentially arbitrary.

Nevertheless, after ordering the connections by score, many renowned hematopoietic regulators appeared in the top ranks, which suggested that the derived pairing scores accurately reflected the biology [↪ section 12.2]. Yet, concerns about the applicability of the HPC-7 cell model to MLL-AF9 leukemia (in particular to the Dnmt1$^{-/chip}$ genotype) as well as the technical resolution limit called for a rigid scrutineering.

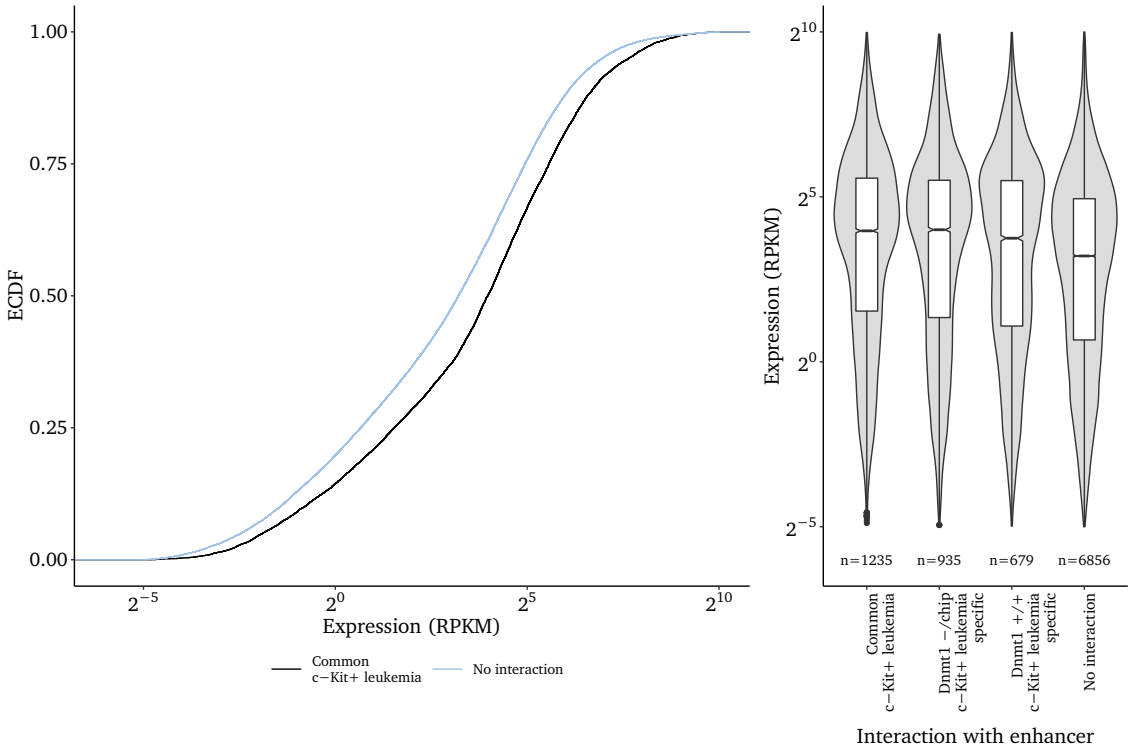


Figure S12.1: Empirical cumulative distribution function of the average expression measured by RNA-seq for transcripts, which are presumably targeted by an enhancer in both genotypes and such, which are not (left panel). To the right, separate categories for genotype-specific enhancers are additionally introduced [▷ Table S 12.1]. Transcripts, whose promoters are targeted by multiple enhancers are included in all applicable categories, thus the sum of all transcripts shown in the plot ($n = 9705$) exceeds the number of expressed transcripts ($n = 9505$).

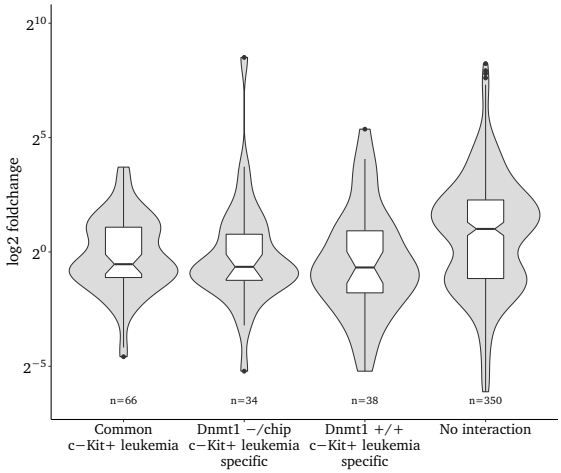
In theory, an enhancer should increase the transcription of a targeted transcript. Thus, we hypothesized that relevant expression differences between genes, which are presumably targeted by an enhancer and such that are not, should exist. Such transcripts, which were presumably targeted by an enhancer showed a significantly higher expression average

Transcript category	Median RPKM	
No interaction	9.33	$<1 \times 10^{-4}$
Dnmt1$^{+/+}$ c-Kit^{+} leukemia specific	13.54	$<1 \times 10^{-4}$
Dnmt1$^{-/chip}$ c-Kit^{+} leukemia specific	16.20	$<1 \times 10^{-4}$
Common c-Kit^{+} leukemia	16.03	

Table S12.1: Median expression of the data shown in Figure S12.1. The brackets to the right indicate the resulting, adjusted p-values for the most relevant contrasts by Tukey's all-pair comparison of linear hypotheses.

across all samples than genes, for which we could not elucidate a potential interaction with a cis-regulatory element (Kruskal-Wallis rank sum test, $p < 2.2 \times 10^{-16}$) [▷ Figure S 12.1, left panel].

Expression differences were smaller for transcripts targeted by common, Dnmt1$^{+/+}$ or Dnmt1$^{-/chip}$-specific putative enhancers [▷ Figure S12.1, right panel]. Although all three categories were significantly higher expressed than the control, Dnmt1$^{+/+}$ was notably diminished relative to the others [▷ Table S12.1]. We also investigated whether expression differences existed between the genotypes, especially for transcripts presumably targeted by genotype-specific enhancers. However, we did not notice any significant differences in overall absolute expression [▷ data not shown].



Transcript category	Median logFC
No interaction	0.99
Dnmt1$^{+/+}$ c-Kit^{+} leukemia specific	−0.68
Dnmt1$^{-/chip}$ c-Kit^{+} leukemia specific	−0.65
Common c-Kit^{+} leukemia	−0.54

Figure S12.2: Visualization and tabular summary of the 488 transcripts, which were significantly differentially expressed in Dnmt1$^{-/chip}$ (3.64 % to 5.60 % depending on category). The distribution of foldchange relative to Dnmt1$^{+/+}$ is plotted.

Next, we focused on the Dnmt1$^{-/chip}$ vs. Dnmt1$^{+/+}$ differentially expressed transcripts within the categories (5.34 %, 3.64 %, 5.60 % and 5.11 % respectively). Since promoter hypomethylation hardly affected expression in Dnmt1$^{-/chip}$ [↪ subsection 8.3.1, p.59], we investigated whether aberrant enhancer recruitment had an influence. In fact, we found differences, but not as expected. Surprisingly, the majority of enhancer-targeted differentially expressed transcripts was downregulated in Dnmt1$^{-/chip}$ c-Kit^{high} cells regardless of the enhancer category [▷ Figure S12.2]. Even the differentially expressed transcripts

targeted by Dnmt1$^{-/chip}$-specific enhancers were predominantly downregulated. The reason for this remained enigmatic, but it should be stressed that all assignments were based on HPC-7 cells and this category likely had the highest error rate of all.

Although these findings warrant confirmation and just very few transcripts changed expression at all, the consistent downregulation of all enhancer categories in Dnmt1$^{-/chip}$ suggest that regulatory chromatin contact might have been perturbed in select cases, possibly by differential methylation [↪ section 11.1, p.89]. Such perturbations likely affect long-range interactions more readily than short-range contacts [328–331], which was in line with our finding that differentially expressed transcripts presumably devoid of (long-range) enhancer interaction tended to be upregulated in Dnmt1$^{-/chip}$ [▷ Figure S12.2].

Admittedly, this interpretation is highly speculative and a simpler explanation relying on fewer assumptions is, that downregulation in these categories was simply favored by the fact that the basal expression was higher in the first place [▷ Figure S12.1].

12.2 Targeted genes and biological implications

Most of the 2750 connections between CAGE-defined enhancers and expressed genes in MLL-AF9 leukemia, which we inferred from the HPC-7 murine blood stem/progenitor cell model [246] linked one gene promoter region[2] to an individual enhancer only. Among the top100 interactions according to the confidence score [↪ Table S14.3, p.134], which reflected the number of connecting reads in Hi-C as well as a possible reference in the FOCS collection of validated pairings [326], we could identify many renowned hematopoietic regulators. On top of that, the majority of involved enhancers originated from strongly accumulated clades (73 %).

12.2.1 Exemplary single enhancer-promoter connections

Irf2bp2: The highest ranked gene was Interferon regulatory factor 2-binding protein 2 (Irf2bp2), the upregulation of which diminishes the induction of apoptosis after doxorubicin treatment [332]. Studies in mouse as well as zebrafish suggest an important role in suppressing erythroid genes and steering myeloid development [333,334]. In addition, a Irf2bp2 fusion can give rise to a subtype of acute promyelocytic leukemia (APL) [335], an involvement in MLL-AF9 is so far unheard of.

Pten: The second placed gene on the list, Phosphatidylinositol 3,4,5-trisphosphate 3-phosphatase and dual-specificity protein phosphatase (Pten), was puzzling for a variety of reasons. First, the interacting enhancer was genotype-specific for Dnmt1$^{-/chip}$, whereas all other top connections referred to common enhancers. Second, the transcript in question was also significantly downregulated for both contrasts: Its expression was lower in leukemic stem cells than blasts (logFC -0.91 for c-Kit^{low} vs. c-Kit^{high}) as well

[2] sometimes comprising several transcripts

as in Dnmt1$^{-/chip}$ (logFC −0.49 for Dnmt1$^{+/+}$ vs $^{-/chip}$). Third, Pten is a dual-specificity phosphatase dephosphorylating proteins as well as lipids. The latter function permits the protein to antagonize the PI3K-AKT/PKB signaling pathway by dephosphorylating phosphoinositides [336] and confers a tumor suppressor role to Pten in MLL-AF9 leukemia [337, 338]. Therefore, a sustained expression of Pten was counterintuitive.

However, in-depth data inspection and literature search in part alleviated the contradiction. Although the determined downregulation in LSCs and Dnmt1$^{-/chip}$ was generally confirmed, a closer inspection of the measured RNA-seq values revealed a high degree of variation among the biological replicates. This was in accordance with the notion, that already modest changes in Pten activity affect cancer susceptibility and progression in mouse models [339]. In solid cancers, haploinsufficiency rather than full inactivation is more frequent [340, 341], since complete loss induces replication stress and senescence [342]. Therefore, the remaining copy is typically retained until aggressive metastatic stages have emerged and an intricate transcriptional [343], post-transcriptional, and post-translational regulation of Pten activity ensures further tumor progression [reviewed in 344, 345]. Taken together, Pten was no improbable candidate in the list, although its role in MLL-AF9 leukemia warrants further investigation.

Fosl2 / Spred1: Further down the list, the next candidates were less puzzling. The Fos-related antigen 2 (Fosl2) dimerizes with Jun to activate LIF transcription and its upregulation is generally associated with therapy resistance in various leukemia subtypes [346]. The next gene, Spred1, is highly expressed in interleukin-3 (IL-3)-dependent hematopoietic cell lines [347] and more than 70 fold upregulated in MLL-AF4 leukemia [348]. Additionally, a predisposition to leukemia in children might be associated with the gene [349, 350].

12.2.2 Selected loci featuring an interplay of multiple enhancers

Many genes implicated in tumorigenesis or hematopoietic regulation were contained in the list of high-confidence enhancer-promoter interactions [↪ Table S14.3, p.134]. None the less, we additionally exploited the characteristic that key regulatory genes often exhibit enhancer redundancy [132] to allow for the selection of even better candidates. Therefore, we ascertained which transcripts ($n = 200$) were targeted by multiple enhancers according to the HPC-7 Hi-C data and ordered the list by the cumulative confidence scores [↪ Table S14.4, p.139]. Subsequently, we examined the transcription factor motifs and clade assignments of the enhancers involved.

Irf2bp2: Aforementioned Interferon regulatory factor 2-binding protein 2 (Irf2bp2), which already led the individual ranking, also dominated the cumulated score. Residing in an approximately 15 kb hypomethlyated region, the gene was expressed in c-Kit^{high} cells with an average of 89.07 RPKM and exhibited no genotype specificity [▷ Figure S12.3].

A total of five enhancers presumably regulated the gene, although interaction strength (at least in HPC-7 cells) clearly set one enhancer apart from the others. The enhancer *chr8*:129118460-

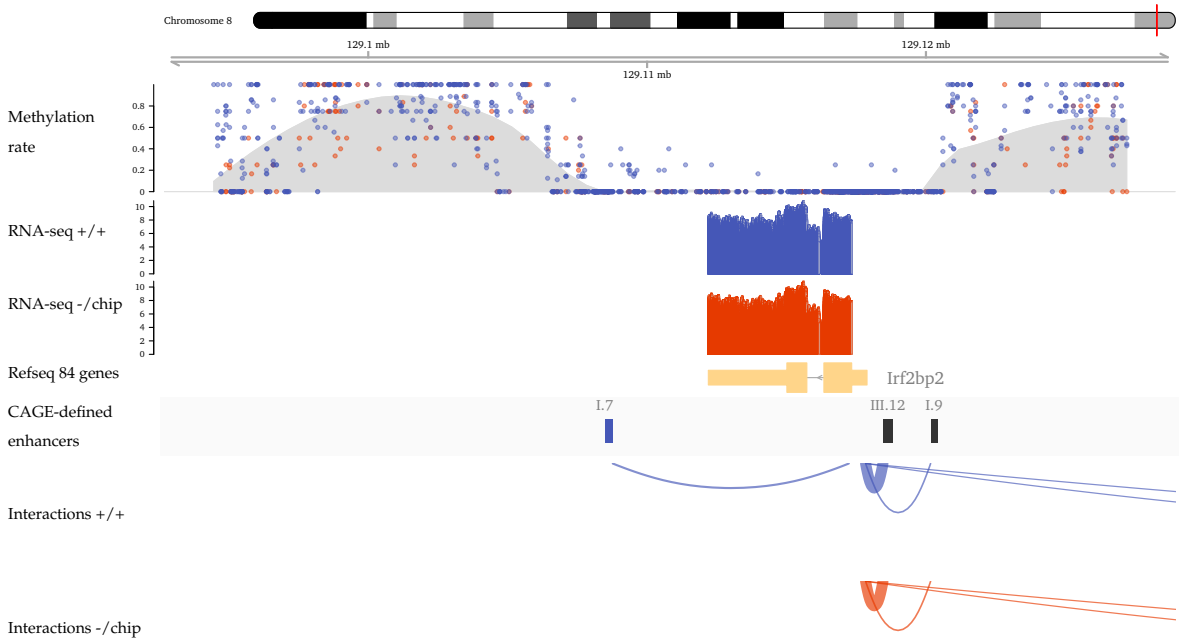


Figure S12.3: Visualization of the genomic region surrounding Irf2bp2. Shown in blue is all data referring to Dnmt1$^{+/+}$, while Dnmt1$^{-/chip}$ is represented in red color. RNA-seq data is log-scaled to base 2 and interaction frequency is conveyed by thickness of the arcs. Methylation rate of single CpGs is given as decimal fraction and displayed as dots, the gray area underneath marks the LOESS smoothed methylation rate. Mind the unusual long 3′-UTR of Irf2bp2 (thin stroke) in contrast to the thickly depicted protein coding sequence (CDS) of the transcript.

129118806, which was assigned to the *Lymphoid + Progenitors* (III) cluster in the strongly accumulated clade *III.12*, exhibited many de novo motifs such as the presumable Mll2-motif ***DeNovo.***SSCGCGGCCTSS in conjunction with recognition sequences of the Early B-cell factor 1 (Ebf1) [351] and one of the Transcriptional enhancer factors (Tead1-Tead4) [▷ Figure S12.4, middle row].

To a lesser extent, also four other enhancers were supposedly involved in the expression of Irf2bp2. Of those, ***chr8:***129108484-129108768 was remarkable, since it was the only genotype-specific enhancer of the gene and only to be detected in Dnmt1$^{+/+}$ [▷ Figure S 12.4, top left panel]. Assigned to the strongly accumulated clade *I.7*, it featured the motifs ***DeNovo.***SSCGCGGCCTSS and ***DeNovo.***CCGGRCGGCG, which suggested it can also be recognized by Mll2. Its interaction frequency score in HPC-7 cells was just about 10 % of that of the main enhancer (12.5 vs. 123.66) but still double the score of the third-frequent enhancer.

chr8:129120168-129120414 - like the main enhancer - was a common cis-regulatory element and positioned technically downstream of the transcript to the distal end of the chromosome, but upstream in relation to the promoter. The most relevant motif of the enhancer was clearly ***PU.1.ThioMac.****PU.1.ChIP.Seq.Homer* [▷ Figure S12.4, top right panel], providing a rationale why the enhancer was regarded as a member of the strongly accumulating clade *I.9* of the *Common* (I) cluster. The confidence score of the interaction was 6.0, which albeit seemingly weak, was still narrowly above Q_3 (above the 75*th* percentile of all confidence

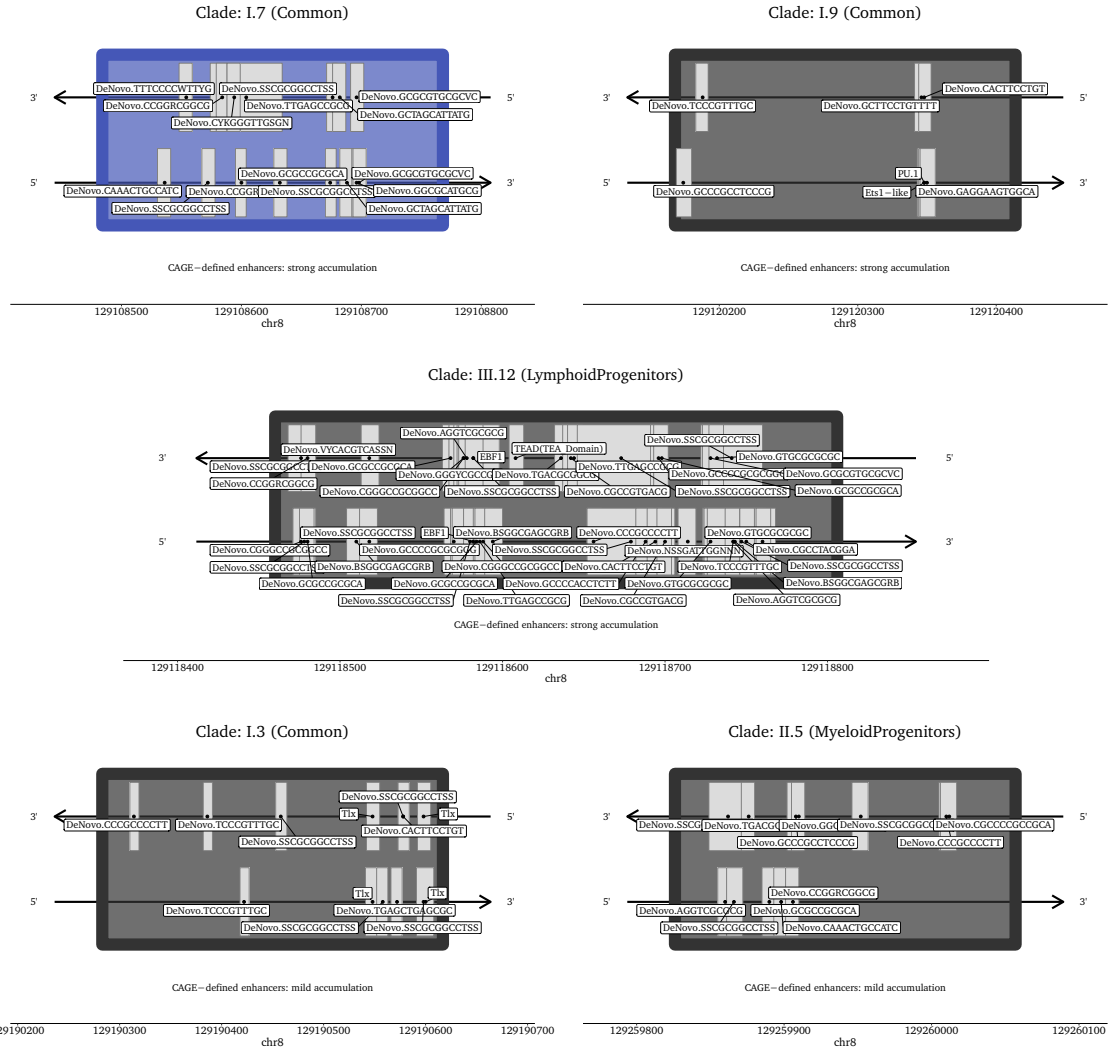


Figure S12.4: The five enhancers presumably targeting Irf2bp2 are depicted with the contained transcription factor biding motifs. Colors indicate genotype-specificity: Common enhancers are shown in black, blue and red denote elements specific to Dnmt1$^{+/+}$ and Dnmt1$^{-/chip}$ respectively.

scores).

The forth and fifth enhancer of Irf2bp2, ***chr8*:**129190283-129190617 and ***chr8*:**129259826-129260057 are not shown in Figure S12.3, as their respective position 72.7 kb and 142.2 kb distal of the transcript could not be drawn at scale. Additionally, their confidence scores of 1.50 and 1.0 made us doubt about a relevant assignment. Yet, no other possible target genes were located nearby and both shared some properties already familiar to us: They were common to the genotypes, belonged to mildly accumulating clades and both featured the motif ***DeNovo.***SSCGCGGCCTSS [▷ Figure S12.4, bottom row]. Taken together, it was tempting to speculate that the intricate regulation observed at the gene's locus implicated Irf2bp2 in MLL-AF9 leukemic transformation or progression.

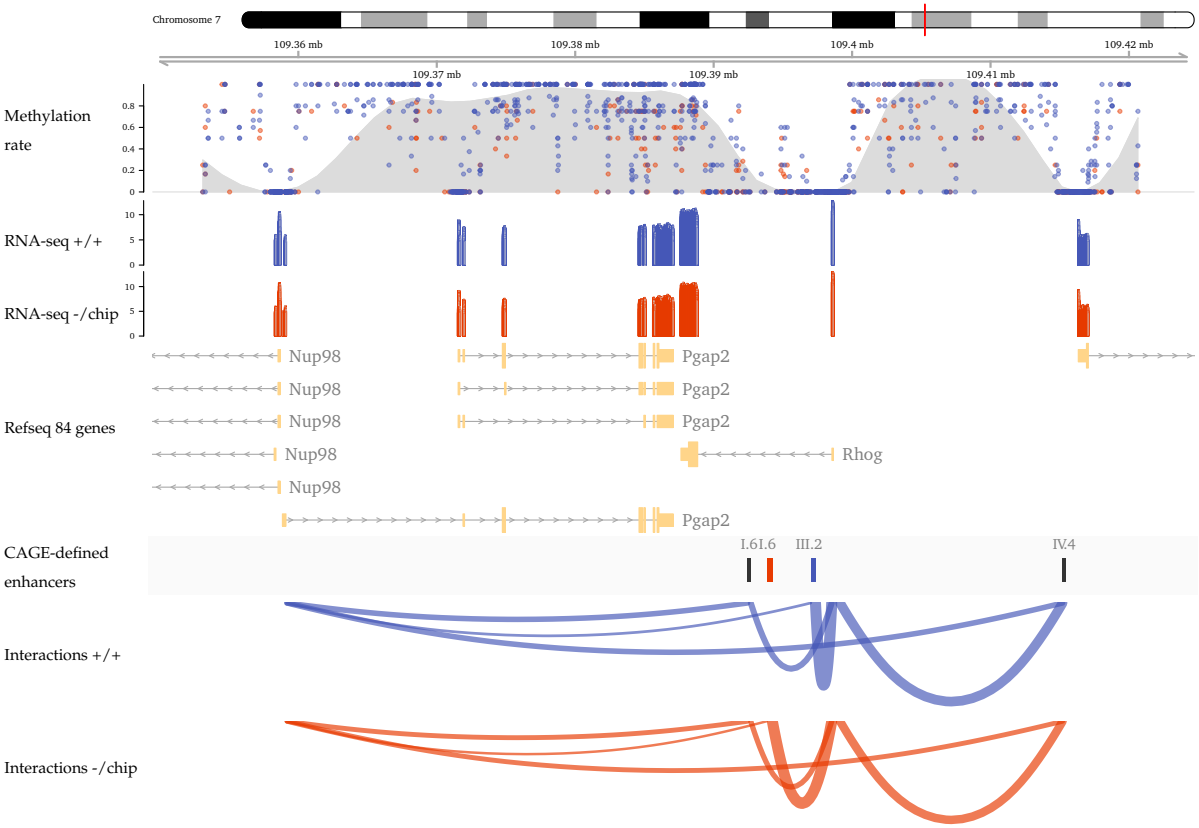


Figure S12.5: Representation of the second ranking gene locus comprising the promoters of Nup98, Pgap2, Rhog and Stim1 (from left to right). The top track contains a scatterplot representation of single CpG methylation rates as well as a LOESS smooth thereof depicted in gray. RNA-seq data is log-scaled to base 2 and (like all other relevant items) colored by genotype: Red is used for data referring to Dnmt1$^{-/chip}$ and blue constitutes Dnmt1$^{+/+}$ items. Thickness of the arcs conveys the frequency of the respective interaction.

Rhog/Nup98: In contrast to Irf2bp2, Ras homolog family member G (Rhog), the second ranking candidate of the cumulative gene list [↪ Table S14.4, p.139], was an unlikelier contender based on single enhancer-promoter interactions as the most relevant interaction was rated with a confidence score of 41.86, the fifteenth highest in total. Nevertheless, the cumulative score of four strong enhancers justified the overall ranking.

The locus was particularly noteworthy due to the complex regulatory circuitry inferred from the HPC-7 Hi-C data. Provided comparability, a total of four enhancers, two of which genotype-specific, targeted three different genes. While Rhog was clearly the main target, all four enhancers had weaker ties to a second fragment comprising the promoters of the Nuclear pore complex protein 98 (Nup98) on the minus strand and a long isoform of the Post-GPI attachment to proteins factor 2 (Pgap2) on the plus strand.

Lacking strand specific RNA-seq, it could not be determined precisely, which of the two genes was targeted primarily, but both genes were expressed and reads in support of the junction in the long Pgap2 isoform could be also identified. However, literature research suggested that Pgap2, which is involved in the lipid remodeling steps of glycosylphosphatidylinositol (GPI) anchors on GPI-anchored proteins, has a negli-

gible importance for acute myeloid leukemia: In AML with myelodysplastic features, glycosylphosphatidylinositol-anchor protein deficiency was associated with genomic instability and leukemic progression [352]. Moreover, as a byproduct of their research, a working group seeking to improve the CRISPR-Cas9 system [353] deleted Pgap2 in the MLL-AF9 rearranged AML cell line MOLM-13 [354] and obtained viable cells. This led us to the conclusion that rather Nup98 or Rhog were the relevant targets at this locus.

In strong support of Nup98 was its known involvement in leukemogenesis. Being part of the nucleoporin family, it normally constitutes a subunit of the nuclear pore complex, which is embedded in the nuclear envelope and facilitates transport of proteins between the cytoplasm and the cell nucleus in eukaryotes [355]. Some of the nucleoporins are able to detach from the pore complex, move to the inner nucleus and alter transcription [356–358]. In 1996, the leukemogenic potential of gene fusions of Nup98 with Hoxa9 was recognized [359, 360] and later emulated in mouse models [361]. Since Hoxa9 itself has a major role in the development of leukemia, it remained unclear to which degree the pathogenicity could be attributed to Nup98, until further leukemogenic fusions with other genes than Hoxa9 were identified [362, 363]. About a decade ago, a study proved the oncogenic potential of Nup98 fusions, in which an extrinsic plant homeodomain (PHD) finger targeted the joint protein to H3K4me3 marked regions of the genome [364]. When mutations in the PHD fingers abrogated H3K4me3 binding, the leukemic transforming capability was lost, since the differentiation-associated polycomb-mediated removal of the mark could no longer be prevented [364]. Later, it was also shown that Nup98 not only blocks removal of H3K4me3, but can also recruit the Wdr82-Set1A/COMPASS complex to mediate deposition of said histone modification [365]. Considering the abnormally strong H3K4me3 signal at genes in MLL-AF9 leukemic stem cells [366] as well as the congruent pattern at enhancers in strongly accumulated clades [↪ section 10.5, p.84], we were intrigued to proclaim Nup98 as the relevant target of those enhancers due to its close ties with H3K4me3.

However, the confidence scores linking the Ras homolog family member G (Rhog) with the enhancers were clearly higher than those for Nup98. The gene is highly expressed in lymphocytes, yet knock-out mice exhibited only mild phenotypic alterations, possibly due to a functional redundancy with other Rac proteins [367]. Later studies linked the protein RhoG with cytoskeletal reorganization and leukocyte trans-endothelial migration in particular [368].

In terms of signaling, several guanine nucleotide exchange factors of the Vav family are capable to mediate the GDP/GTP exchange at RhoG [369]. Intriguingly, RhoG is, jointly with the scaffold protein Elmo2 and the integrin-linked kinase Ilk, part of a tripartite complex and can activate RAC1 [370]. RAC1 activation in turn is known to promote leukemogenesis as well as interactions with the bone marrow microenvironment [371] and to confer chemotherapy resistance to MLL-AF9 leukemic cells [372] . Therefore, expression of the Rhog gene was perfectly sound[3] and might, possibly via the phosphorylation of

[3] also Elmo1/Elmo2 and Ilk were expressed

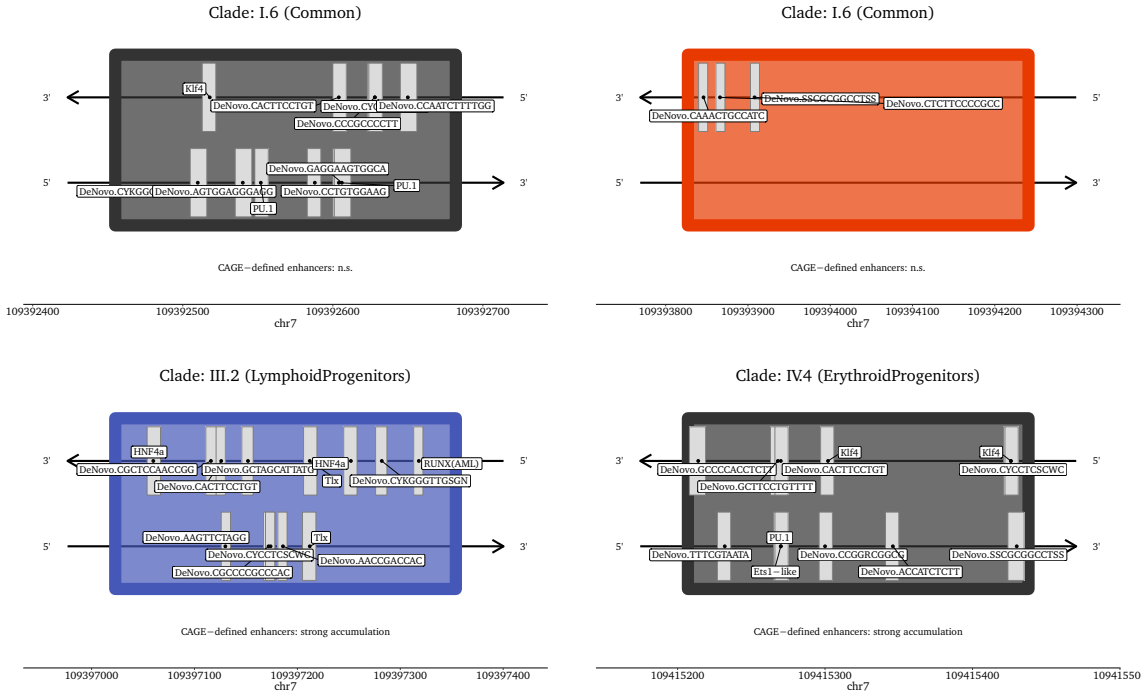


Figure S12.6: Schematic representation of the four enhancers presumably involved in regulating the expression of Rhog and Nup98 in MLL-AF9 leukemia. Genotype-specificity is indicated by the colors red (*-/chip*), blue (*+/+*) and black (common). Gray boxes symbolize approximate positions of transcription factor binding motifs.

Stathmin [reviewed in 373] promote leukemic cell survival.

Taken together, both RhoG and Nup98 probably need to be expressed in MLL-AF9 leukemia, which is ensured by an intricate interplay of four different, possibly redundant enhancers according to the CAGE-seq in MLL-AF9 leukemia and Hi-C data from HPC-7 cells [▷ Figure S12.5]. Three out of the four enhancers were located in the first intron of the Rhog gene, two of which acted in a genotype-specific manner.

The leftmost enhancer ***chr7***:109392455-109392682 (*I.6*) [▷ Figure S12.6, top left] interacted almost equally frequent with RhoG and Nup98 (confidence score of 17.86 and num15.67 respectively), whereas the others clearly favored RhoG over Nup98 (from left to right: 41.86 vs. 5.17, 35.86 vs. 5.25 and 30.00 vs. 15.70). In particular, the two genotype-specific enhancers ***chr7***:109393827-109394241 (*I.6*) [▷ Figure S12.6, top right] and ***chr7***:109397023-109397354 (*III.2*) [▷ Figure S12.6, bottom left] interacted relatively weakly with Nup98, anyhow a confidence score in the range of 5 still corresponded to the 60th percentile of all scores.

With regard to the transcription factor binding sites, the two common enhancers ***chr7***:109392455-109392682 (*I.6*) and ***chr7***:109415207-109415438 (*IV.4*) prominently featured motifs for PU.1 as well as Klf4. Interestingly, decreased expression of PU.1 [129, 130] and Klf4 [374] rather than (over)expression contributes to AML pathogenesis. Indeed, while it is possible to generate induced pluripotent stem (iPS) cells [375] from MLL-AF9 leukemic stem cells (LSCs)

by expressing the *Yamanaka* reprogramming transcription factors[4] artificially, their expression is generally mutually exclusive with sustained MLL-AF9 activity [376]. According to our RNA-seq data, Klf4 was expressed at low levels in LSCs (avg. RPKM = 4.89) and elevated levels in the leukemic bulk (avg. RPKM = 130.70) .

Remarkably, the strongest interacting enhancer (at least for Rhog, confidence score 41.86) was ***chr7***:109393827-109394241 (*I.6*), an enhancer specific to the Dnmt1$^{-/chip}$ genotype. It featured the ***DeNovo.***SSCGCGGCCTSS motif, but hardly any other putative transcription factor binding sites and was also not assigned to a strongly accumulated clade [▷ Figure S 12.6, top right]. This finding adumbrated a multi-layered, complex regulation of relevant genes beyond the enhancers in strongly accumulated clades and Mll2-binding.

Ikzf2: A good example of a gene with relevance for leukemogenesis, which was regulated by a completely different set of motifs and enhancers, was Zinc finger protein Helios (Ikzf2) [▷ Figure S12.7]. The gene was recently identified as crucial mediator of leukemic stem cell self-renewal and differentiation block in AML [377] and was ranked 9th in the combined gene list.

Remarkably, we could detect an unexplained RNA-seq signal upstream of the reference promoter of Ikzf2, which resembled an additional exon. While the mouse reference lacked an according exon, cross-species alignments of the human Ikzf2 recorded a corresponding feature. Deeming a contamination of the library with human cDNA unlikely, we therefore attributed this signal to a yet unannotated transcript variant with an extended 5′-UTR.

Its expression was sustained by two enhancers, the wild-type-specific ***chr1***:69729203-69729392 and the common ***chr1***:69735657-69735883, both affiliated with the *Lymphoid + Progenitors* (III) cluster [▷ Figure S12.8]. The respective clade *III.11* had not accumulated CAGE-defined enhancers or exhibited a particular epigenetic signature [▷ data not shown]. Nevertheless Ikzf2 was consistently expressed in Dnmt1$^{+/+}$ as well as Dnmt1$^{-/chip}$ leukemic cells (avg. RPKM = 31.29).

Except one instance of ***DeNovo.***YAAAAAAAAV, the two enhancers comprised no enriched de novo motifs, which however was to be expected based on the clade assignment. A motif for the Hepatocyte nuclear factor 4α (Hnf4a) was present in the common enhancer, however the corresponding transcription factor was not expressed in MLL-AF9. The wild-type specific enhancer ***chr1***:69729203-69729392 featured motifs for the POU domain, class 2, transcription factor 2 (Pou2f2) and POU domain, class 5, transcription factor 1 (Pou5f1), more commonly known by their legacy names Oct2 and Oct4. While Pou5f1/Oct4 was not active in MLL-AF9 leukemia, Pou2f2/Oct2 was expressed and also reported to have a potential pro-survival function in AML [378]. Thus, Oct2 potentially sustained the expression of Ikzf2 in Dnmt1$^{+/+}$, however how Dnmt1$^{-/chip}$ ensured sufficient levels remained elusive.

[4] Oct4, Sox2, Klf4 and c-Myc

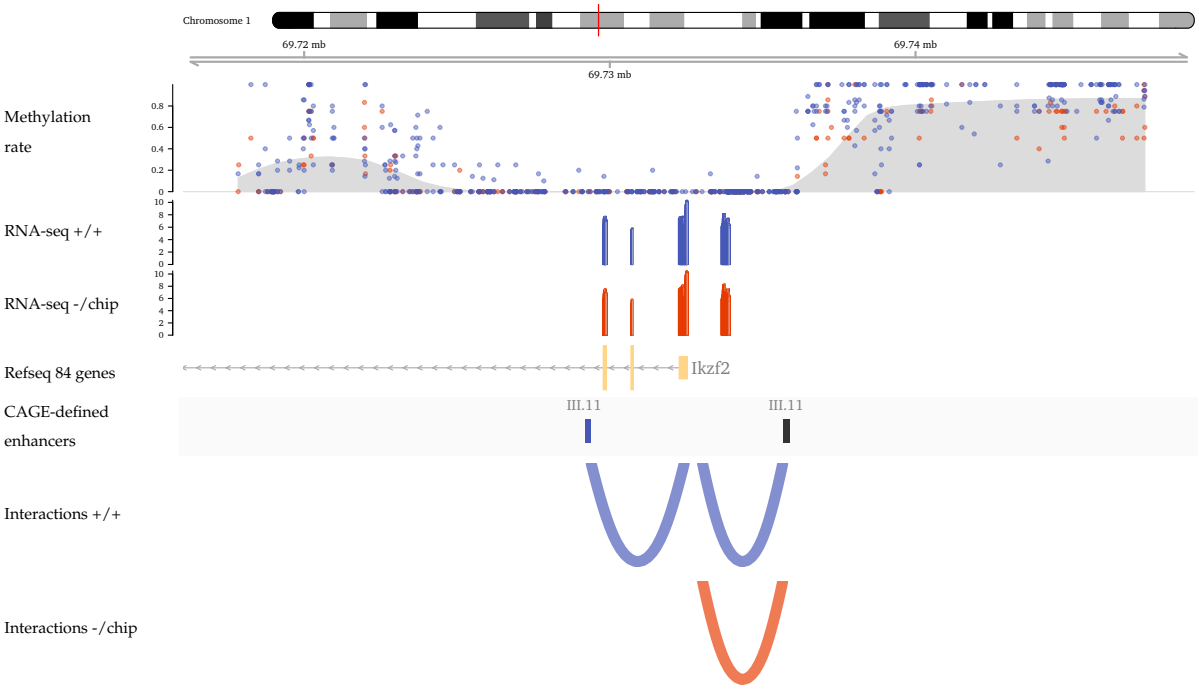


Figure S12.7: Visualization of the Ikzf2 genomic region. RNA-seq data is log-scaled to base 2 and interaction frequency is conveyed by thickness of the arcs. Decimal fractions represent the methylation rate of single CpGs as colored dots, a LOESS smooth shown in gray was applied to determine the local methylation trends. Data colored in blue refers to Dnmt1$^{+/+}$, while Dnmt1$^{-/chip}$ is presented in red color.

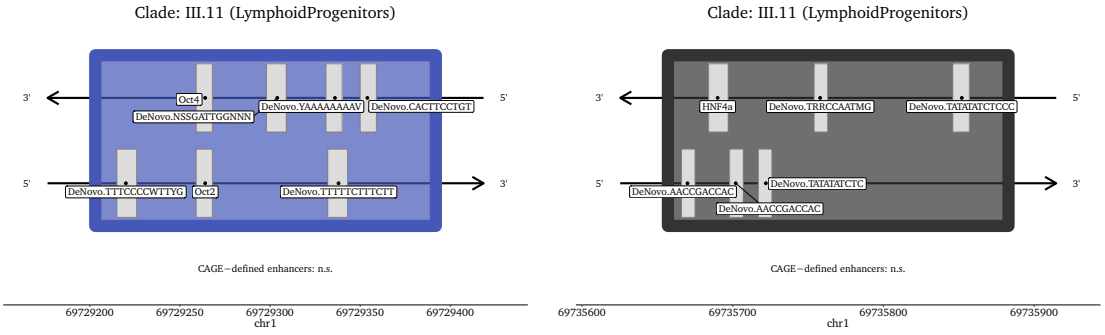


Figure S12.8: Depiction of the two CAGE-defined enhancers assigned to Ikzf2 by HPC-7 Hi-C. Blue and black represent the genotype specificity (Dnmt1$^{+/+}$ and common respectively) and transcription factor binding motifs are conveyed as gray boxes.

12.2.3 Preliminary CRISPR-dCas9 experiments

To test the effects of a particular gene on leukemia progression or self-renewal, we had mostly worked with shRNA-mediated knockdowns or utilized a vector containing the human Ubc-promoter to heterologously express a cDNA copy.

Primarily the candidate genes, which emerged from the integrated methylome and transcriptome analysis [↪ section 8.3, p.59] were experimentally tested in this manner. Depending on, whether our bioinformatic analysis suggested a beneficial up- or downregulation of the gene, we chose the respective method and lentivirally transduced leukemic cells in vitro for constitutive expression of the cDNA or shRNA. About ten genes were tested without success and nixed from the list of potential therapeutic candidates [▷ data not shown].

To test the effects of a particular enhancer, we however required a different experimental approach and opted for a CRISPR-Cas9-based protocol. Hardly any method developed in recent years experienced such rapid acceptance and development as CRISPR-Cas9 did over the last decade. This extended the field of application of the method, which was initially conceived for genome editing and engineering [reviewed in 379], to a whole range of scientific endeavors [reviewed in 380, 381].

Repression through CRISPR interference (CRISPRi) was our preferred choice to test single enhancers in vitro and potentially also in vivo. This method harnesses a nuclease-deficient Cas9 enzyme (dCas9), which is still capable of receiving guidance from single guide RNAs (sgRNA) and can be recruited to specific genomic locations in this manner. Without altering the nucleotide sequence, it may block other DNA-binding complexes such as the mediator complex [68] and thereby silence genes. To strengthen its repression efficiency, the dCas9 used in our protocol was fused to a Krüppel-associated box (KRAB) natively found in a group of repressive zinc-finger proteins [23, 24]. The full construct named *pHR-SFFV-KRAB-dCas9-P2A-mCherry* was cloned in the laboratory of Jonathan Weissman and deposited at Addgene with the accession number 60954 [382]. We also obtained the corresponding sgRNA vector *pU6-sgRNA EF1Alpha-puro-T2A-BFP* (Addgene 60955).

While the two vectors had initially been used for a genome-wide loss-of-function screen [382], we sought to utilize it to target a selection of a few hundred enhancers in parallel for validation purposes. To prepare for this experiment, we cloned about a dozen single sgRNAs as a proof of concept. The test set comprised promoters of published crucial regulators of MLL-AF9, their closest enhancers (regardless of clade enrichment) and a few non-targeting controls [↪ Table S 14.5, p.142]. However, while we could successfully transduce all guide RNAs into MLL-AF9 leukemic cells and establish clones, transduction of *pHR-SFFV-KRAB-dCas9-P2A-mCherry* failed repeatedly. We did not succeed in producing stable dCas9-clones, regardless of whether we had transduced the guides previously or not.

Eventually, we also tried to switch from the transcriptional modulation to a nuclease-

mediated knockout requiring only one plasmid in an alternative approach. Unfortunately, experiments with the completed constructs based on *lentiCRISPR-v2* (Addgene 52961) from Feng Zhang's laboratory [383] were not commenced anymore due to time constraints.

12.3 Assessment of Mll2 target genes

In the previous chapter, it was shown that strongly accumulating clades are characterized by a Mll2-signature. Although Mll2 seemed to be an unlikely candidate, since it had already been subject to some less promising investigations [384], a detailed study from the laboratory of Patricia Ernst highlighted the importance of Mll2 for MLL-AF9 leukemia [385].

In aforesaid publication, RNA-seq was used to characterize the effects of Mll1 or Mll2 deletion in the context of MLL-AF9-transduced leukemia derived from c-Kit^{high} primary bone marrow cells. Analysis of the data provided, amongst others, a list of 171 genes, the expression of which was altered as a result of Mll2 deletion.

However, the authors did not address, why these genes were affected. Having derived a Mll2 binding motif from our enhancers, we were intrigued to see, if this motif would be enriched in the promoters or enhancers of said genes. Therefore, we downloaded the expression data accompanying the study and integrated it with our own data. Since all our data was still aligned to the older reference genome *NCBI37/mm9* instead of *GRCm38/mm10*, we had to realign and reanalyze the published data from scratch utilizing our default RNA-seq pipeline [↪ section 8.2, p.57].

Oddly enough, although the alignment rates to the reference genome *NCBI37/mm9* were extremely good at 95 % to 98 %, only less than half of the aligned reads were found within reference transcripts (46 % to 48 %). This finding prompted us to test our own experimentally derived transcriptome [↪ section 9.1, p.65], however, just 41 % to 43 % of the reads mapped to our experimentally derived transcriptome. Thus, the non-reference transcripts identified in our samples also did not explain these results. We did not follow up on this issue, since it was neither our dataset nor relevant to our research, but it was irritating that this discrepancy was not addressed by the authors [385].

In total, we could identify 8366 expressed reference transcripts[5], of which 1790 (21.4 %) were subject to potential Mll2 regulation. As a proxy for Mll2 recruitment, which was not directly amenable to us, we used the presence of the motif ***DeNovo.***SSCGCGGCCTSS at the promoter (−500 bp to 100 bp around the TSS) or within an assigned enhancer.

Obviously, this lenient approach deliberately overestimated the true Mll2 binding in favor of an inclusive consideration of all possible regulatory action. Despite being useful approximations of transcription factor binding [reviewed in 386], position weight matrix-based methods, like the one used by us, can not accurately predict experimentally deter-

[5] release 84 of the NCBI REFERENCE SEQUENCE DATABASE (REFSEQ) published on September 11, 2017.

mined binding due to oversimplification [387]. For example, in reality, the relationship between binding affinity and probability is not linear [388] and the cell type [389] as well as three-dimensional DNA structure [390] are additional determinants.

Faithful consideration of those factors either by advanced in vitro methods like SMiLE-seq [391] or application of sophisticated new deep learning algorithms [392, 393], such as the convolutional-recurrent neural network FACTORNET [394], was beyond the scope of this project. Therefore, we stuck to the traditional position weight matrix approach (PWM) and scanned promoters and enhancers for presence of the motif ***DeNovo.***SSCGCG-GCCTSS with HOMER.

By this method, we could identify 1790 transcripts (1670 genes) possibly regulated by Mll2 binding. Together with 1038 transcripts, which were presumably regulated by other identified and assigned enhancers, they formed the group of the 2828 apportioned transcripts [▷ Figure S12.9, upper bar]. The regulation of the remainder (5538 transcripts) remained elusive (other promoter motifs were not analyzed) and was irrelevant for the subject in question.

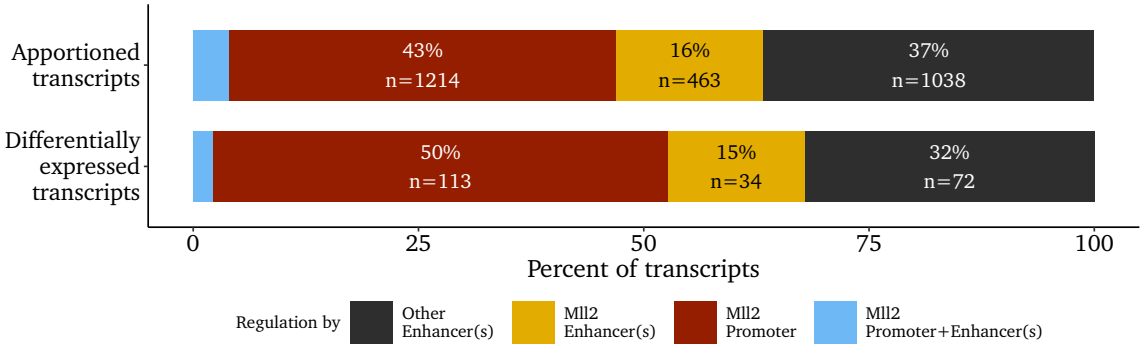


Figure S12.9: Bar graph comparing the regulatory category assignment of transcripts possibly subject to regulation by Mll2 or other CAGE-defined enhancers in MLL-AF9 leukemia (upper bar) to the 211 transcripts differentially expressed upon Mll2 deletion (lower bar).

Interestingly, despite the oversimplified PWM approach, the 211 differentially expressed transcripts (205 genes) reflected the overall assignment quite well [▷ Figure S12.9], which corroborated that Mll proteins preferably bind directly to the promoter of target genes [395]. Yet, we could also identify 34 transcripts (32 genes), which responded presumably due to loss of Mll2 binding at regulating enhancers [↪ Table S14.6, p.143]. Among them was for example Ras homolog family member G (Rhog) [▷ Figure S12.5], which was already described above [↪ subsection 12.2.2].

Approximately two-thirds of the differentially expressed transcripts responded to Mll2 loss by downregulation (151 down, 73 up).

In terms of effect size, the observed expression change (particularly downregulation) was typically more prominent in the promoter category than in the enhancer category [▷ Figure S12.10, pile-up graph to the right]. This finding was likely attributable to potential redundant enhancers and clearly not related to a prior expression bias, since the tran-

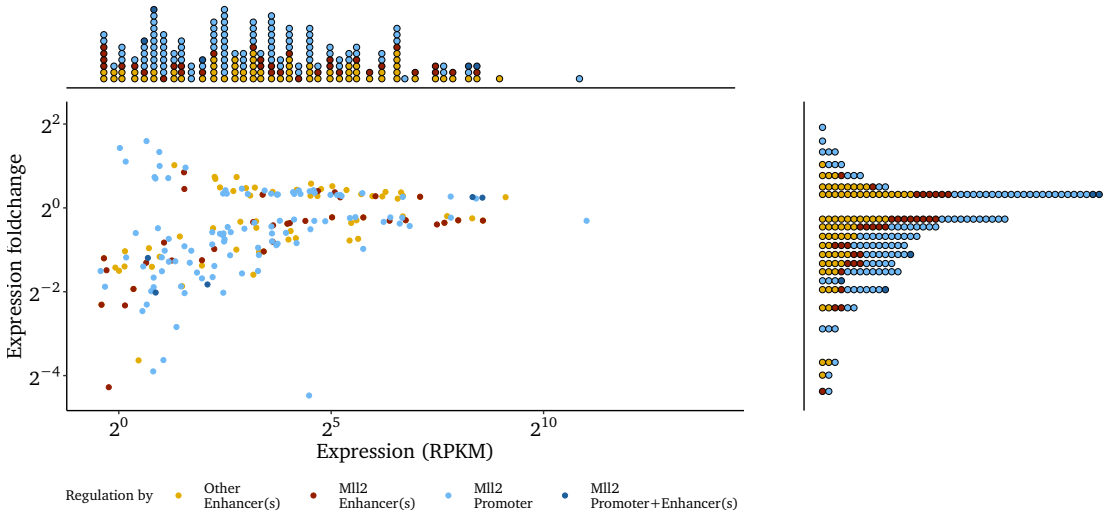


Figure S12.10: Dot plot of the expression pattern after knock-out of Mll2 in MLL-AF9 leukemic cells. Only significantly differentially expressed transcripts are shown. Colors indicate the regulatory assignment. On top and to the right, one-dimensional pile-up plots provide visual aids to assess the absolute number of the significantly differentially expressed transcripts and their respective assignment.

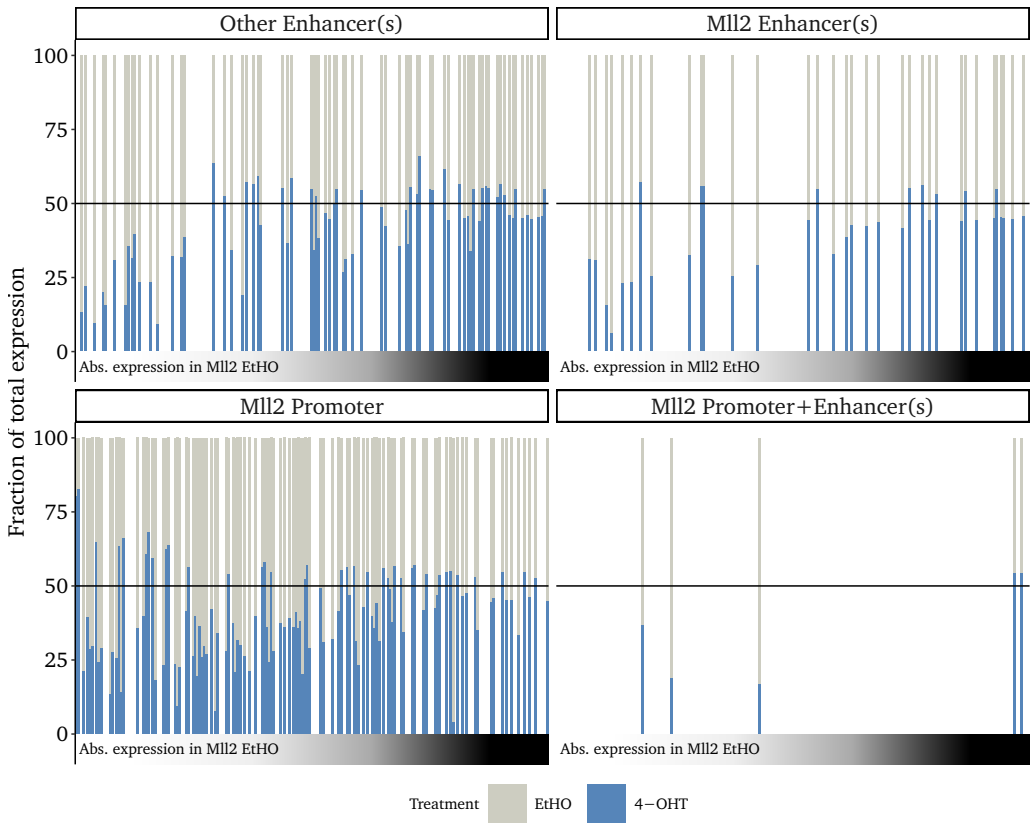


Figure S12.11: Stacked bar plot indicating the transcriptional effect of Mll2 loss after Cre::ER^{T2}-induction by 4-hydroxytamoxifen (4-OHT). Each significantly differentially expressed transcript is depicted as bar scaled to 100 % and ordered by increasing absolute expression in the uninduced ethanol (EtHO) control. Transcripts represented by bars with less than 50 % blue are downregulated following Mll2 loss.

scripts could be found in the full range of the spectrum [▷ Figure S12.11.]

Despite being small in relation to promoter-mediated regulation, both in terms of magnitude and number of affected transcripts, there was a noticeable effect of Mll2-enhancer deficiency. Functionally, some of the respondent genes [↪ Table S 14.6, p.143] were involved in crucial cellular functions such that an effect on self-renewal and leukemogenesis seemed plausible. Nevertheless, none of the candidate genes was experimentally tested anymore.

Thus, it remained elusive, whether the Mll2-signature identified in our CAGE-defined enhancers served a purpose or was a mere passive consequence of Mll2 randomly deviating from promoters. Yet, in murine primordial germ cells, the cis-regulatory binding of Mll2 was already shown to be purposeful [396], which suggested similar mechanisms could apply to MLL-AF9 leukemia and still warrants investigation.

Special emphasis should be placed on DNA methylation at those sites [▷ Figure S11.2, p.92, bottom row], since the CG-rich sequence could easily be subject to EZH2 and PRC2/3-mediated recruitment of DNA methyltransferases [397].

Supplementary Chapter 13

Tables of differentially expressed genes

13.1 RNA-seq Dnmt1$^{-/chip}$ vs. Dnmt1$^{+/+}$

Symbol	Gene name	logFC	logCPM	FDR
Nutm1	Nut Midline Carcinoma, Family Member 1	8.499	4.5619	5.05×10^{-70}
Iqcd	Iq Motif Containing D	4.677	2.1073	4.30×10^{-30}
Myom1	Myomesin 1	−3.541	3.4837	5.28×10^{-16}
2310007B03Rik	Riken Cdna 2310007b03 Gene	5.460	1.2366	1.09×10^{-15}
Dnmt1	Dna Methyltransferase (cytosine-5) 1	−1.010	8.2721	1.09×10^{-15}
Sytl4	Synaptotagmin-like 4	5.518	0.8242	2.77×10^{-14}
Pard6b	Par-6 Family Cell Polarity Regulator Beta	3.028	2.9856	4.27×10^{-11}
Ptgr1	Prostaglandin Reductase 1	−2.465	7.5247	2.08×10^{-10}
Dnajc25	Dnaj Heat Shock Protein Family (hsp40) Member C25	−1.377	4.9307	3.24×10^{-10}
Slain1	Slain Motif Family, Member 1	4.250	2.4108	3.24×10^{-10}
Igtp	Interferon Gamma Induced Gtpase	1.639	4.4597	7.66×10^{-10}
Tsc22d2	Tsc22 Domain Family, Member 2	−1.907	7.1984	7.66×10^{-10}
Unc5cl	Unc-5 Family C-terminal Like	5.033	−0.0536	7.66×10^{-10}
Thbd	Thrombomodulin	−2.266	4.6894	8.39×10^{-10}
Tex19.1	Testis Expressed Gene 19.1	8.230	0.9316	1.00×10^{-9}
Crisp2	Cysteine-rich Secretory Protein 2	3.760	0.4283	1.11×10^{-9}
Shank1	Sh3/ankyrin Domain Gene 1	2.429	1.0887	1.52×10^{-9}
Igsf23	Immunoglobulin Superfamily, Member 23	3.721	0.5758	3.79×10^{-9}
Sycp2	Synaptonemal Complex Protein 2	7.070	3.9029	1.23×10^{-8}
Pde5a	Phosphodiesterase 5a, Cgmp-specific	2.913	−0.1999	3.66×10^{-8}
Best3	Bestrophin 3	7.601	−0.6102	4.33×10^{-8}
Sh2d3c	Sh2 Domain Containing 3c	−1.057	6.1704	4.33×10^{-8}
Adcy6	Adenylate Cyclase 6	2.698	3.3052	7.50×10^{-8}

Table S13.1 continued on next page.

Symbol	Gene name	logFC	logCPM	FDR
Mmp28	Matrix Metallopeptidase 28 (epilysin)	−1.367	4.7735	8.27×10^{-8}
Tbkbp1	Tbk1 Binding Protein 1	−0.851	5.3881	1.70×10^{-7}
Oxct1	3-oxoacid Coa Transferase 1	1.066	6.2966	2.33×10^{-7}
Ccnd2	Cyclin D2	2.903	4.4876	2.67×10^{-7}
Elavl2	Elav (embryonic Lethal, Abnormal Vision, Drosophila)-like 2 (hu Antigen B)	5.001	2.2720	2.87×10^{-7}
Rdx	Radixin	−1.128	8.2798	2.87×10^{-7}
Cdc42bpa	Cdc42 Binding Protein Kinase Alpha	4.135	2.5314	2.88×10^{-7}
Mrgpre	Mas-related Gpr, Member E	4.273	0.7112	2.88×10^{-7}
Crispld2	Cysteine-rich Secretory Protein Lccl Domain Containing 2	−3.156	3.2560	3.77×10^{-7}
Pls3	Plastin 3 (t-isoform)	5.527	0.8269	3.77×10^{-7}
Arid3a	At Rich Interactive Domain 3a (bright-like)	−1.052	6.5944	4.22×10^{-7}
Lgalsl	Lectin, Galactoside Binding-like	1.188	4.3879	6.22×10^{-7}
Gpc1	Glypican 1	−1.305	8.4876	9.16×10^{-7}
Pik3r6	Phosphoinositide-3-kinase, Regulatory Subunit 6	−1.198	5.4375	9.87×10^{-7}
Steap2	Six Transmembrane Epithelial Antigen Of Prostate 2	5.613	2.0955	1.07×10^{-6}
Recql4	Recq Protein-like 4	0.847	5.4634	2.00×10^{-6}
Itga1	Integrin Alpha 1	−2.669	5.2495	2.22×10^{-6}
AI854703	Expressed Sequence Ai854703	3.572	2.8283	2.88×10^{-6}
Ifnlr1	Interferon Lambda Receptor 1	−3.544	1.6826	2.90×10^{-6}
Nov	Nephroblastoma Overexpressed Gene	7.783	4.6535	2.94×10^{-6}
Spin4	Spindlin Family, Member 4	−4.844	0.9814	3.26×10^{-6}
Adcy7	Adenylate Cyclase 7	−0.825	7.6824	9.12×10^{-6}
Cish	Cytokine Inducible Sh2-containing Protein	4.106	4.5748	9.12×10^{-6}
Chsy1	Chondroitin Sulfate Synthase 1	−0.742	6.1695	9.27×10^{-6}
Plekhb1	Pleckstrin Homology Domain Containing, Family B (evectins) Member 1	3.254	3.0294	9.27×10^{-6}
Stap2	Signal Transducing Adaptor Family Member 2	5.247	2.7661	9.27×10^{-6}
Smad6	Smad Family Member 6	−1.983	0.3802	1.12×10^{-5}
Sorbs3	Sorbin And Sh3 Domain Containing 3	3.926	−0.3514	1.36×10^{-5}
Amotl2	Angiomotin-like 2	−2.148	2.9416	1.39×10^{-5}
Abcc8	Atp-binding Cassette, Sub-family C (cftr/mrp), Member 8	−2.689	1.3465	1.57×10^{-5}
Fgl2	Fibrinogen-like Protein 2	2.270	7.8427	1.57×10^{-5}

Table S13.1 continued on next page.

Symbol	Gene name	logFC	logCPM	FDR
Plxnc1	Plexin C1	−1.758	4.1516	1.69×10^{-5}
Stxbp1	Syntaxin Binding Protein 1	1.603	2.9322	2.15×10^{-5}
Galnt7	Udp-n-acetyl-alpha-d-galactosamine: Polypeptide N-acetylgalactosaminyltransferase 7	1.713	6.4291	2.97×10^{-5}
Il2ra	Interleukin 2 Receptor, Alpha Chain	4.841	−0.2558	3.14×10^{-5}
Serpina3g	Serine (or Cysteine) Peptidase Inhibitor, Clade A, Member 3g	3.534	6.5441	3.26×10^{-5}
Irgm1	Immunity-related Gtpase Family M Member 1	0.982	4.7850	3.28×10^{-5}
Fn1	Fibronectin 1	−4.328	7.2518	4.17×10^{-5}
Mov10	Moloney Leukemia Virus 10	1.071	3.9562	4.26×10^{-5}
Itgal	Integrin Alpha L	−1.724	6.2826	4.50×10^{-5}
Kcnk5	Potassium Channel, Subfamily K, Member 5	3.820	0.5188	4.50×10^{-5}
Slc4a5	Solute Carrier Family 4, Sodium Bicarbonate Cotransporter, Member 5	1.856	3.9440	4.50×10^{-5}
Cxxc5	Cxxc Finger 5	3.324	2.8295	4.52×10^{-5}
Arc	Activity Regulated Cytoskeletal-associated Protein	−3.432	1.0333	4.70×10^{-5}
Adgrg1	Adhesion G Protein-coupled Receptor G1	2.102	4.7633	5.00×10^{-5}
H2-T24	Histocompatibility 2, T Region Locus 24	1.051	4.9469	5.01×10^{-5}
Dapk2	Death-associated Protein Kinase 2	−2.483	2.0777	5.11×10^{-5}
Irgm2	Immunity-related Gtpase Family M Member 2	1.289	3.2145	5.11×10^{-5}
Lzts2	Leucine Zipper, Putative Tumor Suppressor 2	2.743	1.5718	5.11×10^{-5}
Aif1l	Allograft Inflammatory Factor 1-like	2.282	−0.4691	5.17×10^{-5}
Slfn5	Schlafen 5	−1.831	3.2923	5.21×10^{-5}
Stap1	Signal Transducing Adaptor Family Member 1	1.485	4.4049	6.49×10^{-5}
Rasal2	Ras Protein Activator Like 2	1.457	3.0174	7.60×10^{-5}
Zfp703	Zinc Finger Protein 703	−0.931	5.2872	7.60×10^{-5}
Ifi47	Interferon Gamma Inducible Protein 47	1.377	5.4407	8.05×10^{-5}
Ikzf3	Ikaros Family Zinc Finger 3	−4.672	1.8174	9.48×10^{-5}
Dnajc6	Dnaj Heat Shock Protein Family (hsp40) Member C6	4.415	4.7046	9.60×10^{-5}
Gbp2b	Guanylate Binding Protein 2b	1.967	2.4760	1.11×10^{-4}
Ltbp3	Latent Transforming Growth Factor Beta Binding Protein 3	2.116	0.2442	1.11×10^{-4}
Plekhg4	Pleckstrin Homology Domain Containing, Family G (with Rhogef Domain) Member 4	3.702	4.4308	1.11×10^{-4}
Il6ra	Interleukin 6 Receptor, Alpha	−1.014	6.0252	1.22×10^{-4}
Tmem108	Transmembrane Protein 108	−3.447	−0.1613	1.34×10^{-4}

Table S13.1 continued on next page.

Symbol	Gene name	logFC	logCPM	FDR
Gal3st3	Galactose-3-o-sulfotransferase 3	4.075	−0.1652	1.35×10^{-4}
Lhx2	Lim Homeobox Protein 2	−4.499	2.9485	1.50×10^{-4}
Pde3b	Phosphodiesterase 3b, Cgmp-inhibited	−2.133	6.5592	1.50×10^{-4}
Bcat1	Branched Chain Aminotransferase 1, Cytosolic	3.318	3.1391	1.54×10^{-4}
Homer2	Homer Scaffolding Protein 2	2.557	−0.9905	1.54×10^{-4}
Cped1	Cadherin-like And Pc-esterase Domain Containing 1	−2.941	0.8620	1.66×10^{-4}
Noxred1	Nadp+ Dependent Oxidoreductase Domain Containing 1	1.789	2.1543	1.66×10^{-4}
Wdr27	Wd Repeat Domain 27	2.371	−0.1211	1.66×10^{-4}
Rnf144b	Ring Finger Protein 144b	−1.824	5.3393	1.69×10^{-4}
4930550L24Rik	Riken Cdna 4930550l24 Gene	7.910	4.5719	1.72×10^{-4}
Myct1	Myc Target 1	5.364	1.6184	1.78×10^{-4}
Parp12	Poly (adp-ribose) Polymerase Family, Member 12	1.562	4.2659	1.89×10^{-4}
Cp	Ceruloplasmin	2.656	3.0930	1.91×10^{-4}
Ica1	Islet Cell Autoantigen 1	2.873	3.6960	1.95×10^{-4}
C1rl	Complement Component 1, R Subcomponent-like	−0.954	5.5941	2.22×10^{-4}

Table truncated. Top 100 genes of 730 shown.

Table S13.1: Differentially expressed genes in RNA-seq for Dnmt1$^{-/chip}$ vs. Dnmt1$^{+/+}$ contrast. Rows are ordered by false discovery rate (FDR).

13.2 Altered KEGG pathways in Dnmt1$^{-/chip}$

KeggID	Pathway	FDR.path	Symbol	logFC	logCPM	FDR.gene
mmu00340	Histidine metabolism	0.003	Aldh3a2	−0.548	5.9958	1.01×10^{-9}
mmu00340	Histidine metabolism	0.003	Aldh3b1	−0.942	6.5499	4.53×10^{-4}
mmu00340	Histidine metabolism	0.003	Aldh7a1	1.260	2.8333	5.58×10^{-3}
mmu00340	Histidine metabolism	0.003	Amdhd1	1.571	0.3903	7.70×10^{-4}
mmu00340	Histidine metabolism	0.003	Hnmt	0.925	4.5221	7.90×10^{-3}
mmu00340	Histidine metabolism	0.003	Maoa	1.564	3.2862	2.55×10^{-4}
mmu04510	Focal adhesion	0.012	Cav2	−1.711	1.0819	8.76×10^{-6}
mmu04510	Focal adhesion	0.012	Ccnd2	2.903	4.4876	8.22×10^{-10}
mmu04510	Focal adhesion	0.012	Flna	−0.682	10.5382	1.91×10^{-9}
mmu04510	Focal adhesion	0.012	Fn1	−4.328	7.2518	1.88×10^{-15}
mmu04510	Focal adhesion	0.012	Itga1	−2.669	5.2495	2.05×10^{-13}

Table S13.2 continued on next page.

KeggID	Pathway	FDR.path	Symbol	logFC	logCPM	FDR.gene
mmu04510	Focal adhesion	0.012	Itgb5	−1.456	5.1769	5.80×10^{-4}
mmu04510	Focal adhesion	0.012	Itgb7	−2.099	6.7239	3.85×10^{-19}
mmu04510	Focal adhesion	0.012	Lama5	2.680	0.7515	2.76×10^{-4}
mmu04510	Focal adhesion	0.012	Lamb2	1.454	1.0083	1.70×10^{-2}
mmu04510	Focal adhesion	0.012	Pak6	2.473	1.1946	1.30×10^{-2}
mmu04510	Focal adhesion	0.012	Prkcb	−0.597	7.5584	1.10×10^{-7}
mmu04510	Focal adhesion	0.012	Prkcg	−1.992	1.8551	6.93×10^{-14}
mmu04510	Focal adhesion	0.012	Pten	−0.507	7.6192	3.93×10^{-9}
mmu04510	Focal adhesion	0.012	Spp1	−2.310	4.0073	7.25×10^{-9}
mmu04510	Focal adhesion	0.012	Tln2	−2.143	4.6763	5.76×10^{-5}
mmu04510	Focal adhesion	0.012	Vav1	−0.469	8.2507	1.22×10^{-3}
mmu04510	Focal adhesion	0.012	Vegfb	0.738	3.7920	1.91×10^{-4}
mmu04810	Actin cytoskeleton	0.021	Cxcl12	−3.164	0.0164	4.15×10^{-3}
mmu04810	Actin cytoskeleton	0.021	Fn1	−4.328	7.2518	1.88×10^{-15}
mmu04810	Actin cytoskeleton	0.021	Itga1	−2.669	5.2495	2.05×10^{-13}
mmu04810	Actin cytoskeleton	0.021	Itgal	−1.724	6.2826	5.79×10^{-15}
mmu04810	Actin cytoskeleton	0.021	Itgb5	−1.456	5.1769	5.80×10^{-4}
mmu04810	Actin cytoskeleton	0.021	Itgb7	−2.099	6.7239	3.85×10^{-19}
mmu04810	Actin cytoskeleton	0.021	Lpar1	−1.992	2.2449	3.21×10^{-4}
mmu04810	Actin cytoskeleton	0.021	Myh10	−0.706	2.4641	5.19×10^{-4}
mmu04810	Actin cytoskeleton	0.021	Myh9	−0.680	9.6629	6.29×10^{-6}
mmu04810	Actin cytoskeleton	0.021	Nckap1l	−0.352	8.6219	3.21×10^{-9}
mmu04810	Actin cytoskeleton	0.021	Pak6	2.473	1.1946	1.30×10^{-2}
mmu04810	Actin cytoskeleton	0.021	Pip5k1b	1.539	6.0386	3.26×10^{-4}
mmu04810	Actin cytoskeleton	0.021	Rdx	−1.128	8.2798	1.47×10^{-8}
mmu04810	Actin cytoskeleton	0.021	Tiam1	−1.023	6.6414	2.79×10^{-2}
mmu04810	Actin cytoskeleton	0.021	Vav1	−0.469	8.2507	1.22×10^{-3}
mmu04810	Actin cytoskeleton	0.021	Was	−0.448	7.4168	1.80×10^{-3}
mmu04810	Actin cytoskeleton	0.021	Wasf2	−0.466	7.5289	5.38×10^{-6}
mmu04925	Aldosterone synthesis	0.039	Adcy6	2.698	3.3052	5.71×10^{-9}
mmu04925	Aldosterone synthesis	0.039	Adcy7	−0.825	7.6824	1.55×10^{-14}
mmu04925	Aldosterone synthesis	0.039	Atp1b1	−3.021	4.5838	5.64×10^{-4}
mmu04925	Aldosterone synthesis	0.039	Camk2g	−0.373	5.9382	1.22×10^{-2}
mmu04925	Aldosterone synthesis	0.039	Cyp11a1	4.557	0.7890	1.20×10^{-3}
mmu04925	Aldosterone synthesis	0.039	Nr4a2	−1.330	3.6135	1.54×10^{-7}
mmu04925	Aldosterone synthesis	0.039	Prkcb	−0.597	7.5584	1.10×10^{-7}

Table S13.2 continued on next page.

KeggID	Pathway	FDR.path	Symbol	logFC	logCPM	FDR.gene
mmu04925	Aldosterone synthesis	0.039	Prkcg	−1.992	1.8551	6.93×10^{-14}
mmu04925	Aldosterone synthesis	0.039	Prkd3	−1.056	6.8645	4.26×10^{-7}
mmu04210	Apoptosis	0.039	Ctsb	−0.614	9.2536	3.49×10^{-10}
mmu04210	Apoptosis	0.039	Ctsh	−2.198	6.8248	2.51×10^{-9}
mmu04210	Apoptosis	0.039	Ctso	0.569	4.6367	1.32×10^{-3}
mmu04210	Apoptosis	0.039	Ctsw	2.406	0.2563	6.21×10^{-5}
mmu04210	Apoptosis	0.039	Fos	−1.268	9.4901	3.31×10^{-8}
mmu04210	Apoptosis	0.039	Gadd45a	−0.856	4.8913	2.27×10^{-4}
mmu04210	Apoptosis	0.039	Gadd45g	−1.079	3.0145	8.81×10^{-14}
mmu04210	Apoptosis	0.039	Gzmb	5.991	−0.2528	3.46×10^{-3}
mmu04210	Apoptosis	0.039	Il3	7.299	1.1247	9.44×10^{-3}
mmu04210	Apoptosis	0.039	Ptpn13	3.764	2.4381	1.44×10^{-3}
mmu04210	Apoptosis	0.039	Tnfsf10	1.138	1.7466	2.49×10^{-2}
mmu04650	NK cell mediated cytotoxicity	0.039	Gzmb	5.991	−0.2528	3.46×10^{-3}
mmu04650	NK cell mediated cytotoxicity	0.039	H2-T23	0.724	5.7598	3.67×10^{-10}
mmu04650	NK cell mediated cytotoxicity	0.039	Ifngr2	−0.464	5.5491	5.04×10^{-19}
mmu04650	NK cell mediated cytotoxicity	0.039	Itgal	−1.724	6.2826	5.79×10^{-15}
mmu04650	NK cell mediated cytotoxicity	0.039	Lck	1.300	0.3347	1.89×10^{-3}
mmu04650	NK cell mediated cytotoxicity	0.039	Prkcb	−0.597	7.5584	1.10×10^{-7}
mmu04650	NK cell mediated cytotoxicity	0.039	Prkcg	−1.992	1.8551	6.93×10^{-14}
mmu04650	NK cell mediated cytotoxicity	0.039	Sh3bp2	−0.639	5.6326	4.94×10^{-12}
mmu04650	NK cell mediated cytotoxicity	0.039	Tnfsf10	1.138	1.7466	2.49×10^{-2}
mmu04650	NK cell mediated cytotoxicity	0.039	Vav1	−0.469	8.2507	1.22×10^{-3}
mmu04621	NOD-like receptor pathway	0.039	Ccl2	−3.259	1.5831	1.12×10^{-25}
mmu04621	NOD-like receptor pathway	0.039	Ctsb	−0.614	9.2536	3.49×10^{-10}
mmu04621	NOD-like receptor pathway	0.039	Gbp2	1.953	2.6229	1.26×10^{-4}
mmu04621	NOD-like receptor pathway	0.039	Gbp2b	1.967	2.4760	5.24×10^{-7}
mmu04621	NOD-like receptor pathway	0.039	Gbp5	2.149	2.4376	3.30×10^{-4}
mmu04621	NOD-like receptor pathway	0.039	Gbp7	1.644	5.6202	4.67×10^{-3}
mmu04621	NOD-like receptor pathway	0.039	Naip2	−0.523	6.7307	9.08×10^{-4}
mmu04621	NOD-like receptor pathway	0.039	Naip6	−0.711	4.1317	7.72×10^{-11}
mmu04621	NOD-like receptor pathway	0.039	Nlrp1b	−0.804	4.5280	8.61×10^{-7}
mmu04621	NOD-like receptor pathway	0.039	Nlrx1	−0.821	5.2469	3.31×10^{-3}
mmu04621	NOD-like receptor pathway	0.039	P2rx7	−2.097	3.9744	4.68×10^{-20}
mmu04621	NOD-like receptor pathway	0.039	Panx1	0.630	4.5045	7.95×10^{-3}
mmu04621	NOD-like receptor pathway	0.039	Trip6	−0.737	4.2466	2.83×10^{-14}

Table S13.2 continued on next page.

KeggID	Pathway	FDR.path	Symbol	logFC	logCPM	FDR.gene
mmu04015	Rap1 signaling pathway	0.039	Adcy6	2.698	3.3052	5.71×10^{-9}
mmu04015	Rap1 signaling pathway	0.039	Adcy7	−0.825	7.6824	1.55×10^{-14}
mmu04015	Rap1 signaling pathway	0.039	Angpt1	2.404	5.1631	2.65×10^{-3}
mmu04015	Rap1 signaling pathway	0.039	Itgal	−1.724	6.2826	5.79×10^{-15}
mmu04015	Rap1 signaling pathway	0.039	Lpar1	−1.992	2.2449	3.21×10^{-4}
mmu04015	Rap1 signaling pathway	0.039	Pard3	1.891	0.6172	1.38×10^{-4}
mmu04015	Rap1 signaling pathway	0.039	Pard6b	3.028	2.9856	7.00×10^{-12}
mmu04015	Rap1 signaling pathway	0.039	Prkcb	−0.597	7.5584	1.10×10^{-7}
mmu04015	Rap1 signaling pathway	0.039	Prkcg	−1.992	1.8551	6.93×10^{-14}
mmu04015	Rap1 signaling pathway	0.039	Prkcz	−2.292	1.1243	2.20×10^{-2}
mmu04015	Rap1 signaling pathway	0.039	Prkd3	−1.056	6.8645	4.26×10^{-7}
mmu04015	Rap1 signaling pathway	0.039	Rapgef2	1.048	5.8919	6.44×10^{-3}
mmu04015	Rap1 signaling pathway	0.039	Tiam1	−1.023	6.6414	2.79×10^{-2}
mmu04015	Rap1 signaling pathway	0.039	Tln2	−2.143	4.6763	5.76×10^{-5}
mmu04015	Rap1 signaling pathway	0.039	Vegfb	0.738	3.7920	1.91×10^{-4}
mmu05323	Rheumatoid arthritis	0.039	Angpt1	2.404	5.1631	2.65×10^{-3}
mmu05323	Rheumatoid arthritis	0.039	Atp6v1b2	−0.638	8.1558	2.27×10^{-7}
mmu05323	Rheumatoid arthritis	0.039	Ccl2	−3.259	1.5831	1.12×10^{-25}
mmu05323	Rheumatoid arthritis	0.039	Ccl3	−1.376	6.6005	9.61×10^{-26}
mmu05323	Rheumatoid arthritis	0.039	Cd28	−2.598	2.0296	3.20×10^{-7}
mmu05323	Rheumatoid arthritis	0.039	Cxcl12	−3.164	0.0164	4.15×10^{-3}
mmu05323	Rheumatoid arthritis	0.039	Fos	−1.268	9.4901	3.31×10^{-8}
mmu05323	Rheumatoid arthritis	0.039	Itgal	−1.724	6.2826	5.79×10^{-15}
mmu04530	Tight junction	0.039	Amotl1	−1.718	1.1191	4.31×10^{-4}
mmu04530	Tight junction	0.039	Amotl2	−2.148	2.9416	3.77×10^{-6}
mmu04530	Tight junction	0.039	Cldn1	−4.738	−0.5478	1.33×10^{-3}
mmu04530	Tight junction	0.039	Cldn15	−1.256	5.4997	7.14×10^{-4}
mmu04530	Tight junction	0.039	Myh10	−0.706	2.4641	5.19×10^{-4}
mmu04530	Tight junction	0.039	Myh9	−0.680	9.6629	6.29×10^{-6}
mmu04530	Tight junction	0.039	Pard3	1.891	0.6172	1.38×10^{-4}
mmu04530	Tight junction	0.039	Pard6b	3.028	2.9856	7.00×10^{-12}
mmu04530	Tight junction	0.039	Prkcz	−2.292	1.1243	2.20×10^{-2}
mmu04530	Tight junction	0.039	Rapgef2	1.048	5.8919	6.44×10^{-3}
mmu04530	Tight junction	0.039	Rdx	−1.128	8.2798	1.47×10^{-8}
mmu04530	Tight junction	0.039	Tiam1	−1.023	6.6414	2.79×10^{-2}
mmu04530	Tight junction	0.039	Was	−0.448	7.4168	1.80×10^{-3}

Table S13.2 continued on next page.

KeggID	Pathway	FDR.path	Symbol	logFC	logCPM	FDR.gene
mmu04062	Chemokine signaling pathway	0.045	Adcy6	2.698	3.3052	5.71×10^{-9}
mmu04062	Chemokine signaling pathway	0.045	Adcy7	−0.825	7.6824	1.55×10^{-14}
mmu04062	Chemokine signaling pathway	0.045	Ccl2	−3.259	1.5831	1.12×10^{-25}
mmu04062	Chemokine signaling pathway	0.045	Ccl3	−1.376	6.6005	9.61×10^{-26}
mmu04062	Chemokine signaling pathway	0.045	Cxcl12	−3.164	0.0164	4.15×10^{-3}
mmu04062	Chemokine signaling pathway	0.045	Cxcl9	2.836	1.0967	1.71×10^{-4}
mmu04062	Chemokine signaling pathway	0.045	Gng10	−0.713	6.4453	2.88×10^{-3}
mmu04062	Chemokine signaling pathway	0.045	Pard3	1.891	0.6172	1.38×10^{-4}
mmu04062	Chemokine signaling pathway	0.045	Pik3r6	−1.198	5.4375	2.43×10^{-39}
mmu04062	Chemokine signaling pathway	0.045	Prkcb	−0.597	7.5584	1.10×10^{-7}
mmu04062	Chemokine signaling pathway	0.045	Prkcz	−2.292	1.1243	2.20×10^{-2}
mmu04062	Chemokine signaling pathway	0.045	Tiam1	−1.023	6.6414	2.79×10^{-2}
mmu04062	Chemokine signaling pathway	0.045	Vav1	−0.469	8.2507	1.22×10^{-3}
mmu04062	Chemokine signaling pathway	0.045	Was	−0.448	7.4168	1.80×10^{-3}
mmu04151	PI3K-Akt signaling pathway	0.048	Angpt1	2.404	5.1631	2.65×10^{-3}
mmu04151	PI3K-Akt signaling pathway	0.048	Ccnd2	2.903	4.4876	8.22×10^{-10}
mmu04151	PI3K-Akt signaling pathway	0.048	Fn1	−4.328	7.2518	1.88×10^{-15}
mmu04151	PI3K-Akt signaling pathway	0.048	Ghr	3.564	0.8257	1.63×10^{-3}
mmu04151	PI3K-Akt signaling pathway	0.048	Gng10	−0.713	6.4453	2.88×10^{-3}
mmu04151	PI3K-Akt signaling pathway	0.048	Il2ra	4.841	−0.2558	4.96×10^{-7}
mmu04151	PI3K-Akt signaling pathway	0.048	Il3	7.299	1.1247	9.44×10^{-3}
mmu04151	PI3K-Akt signaling pathway	0.048	Il6ra	−1.014	6.0252	4.01×10^{-29}
mmu04151	PI3K-Akt signaling pathway	0.048	Itga1	−2.669	5.2495	2.05×10^{-13}
mmu04151	PI3K-Akt signaling pathway	0.048	Itgb5	−1.456	5.1769	5.80×10^{-4}
mmu04151	PI3K-Akt signaling pathway	0.048	Itgb7	−2.099	6.7239	3.85×10^{-19}
mmu04151	PI3K-Akt signaling pathway	0.048	Lama5	2.680	0.7515	2.76×10^{-4}
mmu04151	PI3K-Akt signaling pathway	0.048	Lamb2	1.454	1.0083	1.70×10^{-2}
mmu04151	PI3K-Akt signaling pathway	0.048	Lpar1	−1.992	2.2449	3.21×10^{-4}
mmu04151	PI3K-Akt signaling pathway	0.048	Pik3r6	−1.198	5.4375	2.43×10^{-39}
mmu04151	PI3K-Akt signaling pathway	0.048	Ppp2r5b	0.703	4.2048	2.08×10^{-4}
mmu04151	PI3K-Akt signaling pathway	0.048	Pten	−0.507	7.6192	3.93×10^{-9}
mmu04151	PI3K-Akt signaling pathway	0.048	Sgk1	−1.155	4.9817	1.03×10^{-16}
mmu04151	PI3K-Akt signaling pathway	0.048	Sgk3	−0.915	7.2389	1.44×10^{-2}
mmu04151	PI3K-Akt signaling pathway	0.048	Spp1	−2.310	4.0073	7.25×10^{-9}
mmu04151	PI3K-Akt signaling pathway	0.048	Vegfb	0.738	3.7920	1.91×10^{-4}
mmu04020	Calcium signaling pathway	0.048	Adcy7	−0.825	7.6824	1.55×10^{-14}

Table S13.2 continued on next page.

KeggID	Pathway	FDR.path	Symbol	logFC	logCPM	FDR.gene
mmu04020	Calcium signaling pathway	0.048	Adrb2	−1.559	4.9783	3.52×10^{-13}
mmu04020	Calcium signaling pathway	0.048	Cacna1e	−2.784	1.2865	4.31×10^{-10}
mmu04020	Calcium signaling pathway	0.048	Camk2g	−0.373	5.9382	1.22×10^{-2}
mmu04020	Calcium signaling pathway	0.048	Nos1	3.071	1.3681	4.23×10^{-5}
mmu04020	Calcium signaling pathway	0.048	Orai2	−0.471	5.9257	3.13×10^{-2}
mmu04020	Calcium signaling pathway	0.048	P2rx1	−1.114	5.0075	2.24×10^{-2}
mmu04020	Calcium signaling pathway	0.048	P2rx4	−0.849	4.4766	3.45×10^{-8}
mmu04020	Calcium signaling pathway	0.048	P2rx7	−2.097	3.9744	4.68×10^{-20}
mmu04020	Calcium signaling pathway	0.048	Pde1b	−0.976	7.1071	3.94×10^{-12}
mmu04020	Calcium signaling pathway	0.048	Prkcb	−0.597	7.5584	1.10×10^{-7}
mmu04020	Calcium signaling pathway	0.048	Prkcg	−1.992	1.8551	6.93×10^{-14}
mmu04020	Calcium signaling pathway	0.048	Ptger1	−0.712	5.4770	1.86×10^{-3}
mmu04630	JAK-STAT signaling pathway	0.048	Ccnd2	2.903	4.4876	8.22×10^{-10}
mmu04630	JAK-STAT signaling pathway	0.048	Cish	4.106	4.5748	1.04×10^{-6}
mmu04630	JAK-STAT signaling pathway	0.048	Ctf1	1.658	−0.2859	1.52×10^{-5}
mmu04630	JAK-STAT signaling pathway	0.048	Ghr	3.564	0.8257	1.63×10^{-3}
mmu04630	JAK-STAT signaling pathway	0.048	Ifngr2	−0.464	5.5491	5.04×10^{-19}
mmu04630	JAK-STAT signaling pathway	0.048	Ifnlr1	−3.544	1.6826	1.19×10^{-7}
mmu04630	JAK-STAT signaling pathway	0.048	Il2ra	4.841	−0.2558	4.96×10^{-7}
mmu04630	JAK-STAT signaling pathway	0.048	Il3	7.299	1.1247	9.44×10^{-3}
mmu04630	JAK-STAT signaling pathway	0.048	Il5ra	4.924	2.5166	1.65×10^{-2}
mmu04630	JAK-STAT signaling pathway	0.048	Il6ra	−1.014	6.0252	4.01×10^{-29}
mmu04630	JAK-STAT signaling pathway	0.048	Il6st	−1.170	5.3835	7.27×10^{-17}
mmu04630	JAK-STAT signaling pathway	0.048	Lif	−1.813	1.2655	5.10×10^{-4}

Table S13.2: Genes assigned to significantly enriched KEGG pathways for Dnmt1$^{-/chip}$ vs. Dnmt1$^{+/+}$ contrast (in RNA-seq, $n = 14$). Rows are ordered by FDR of the respective pathway and the genes within alphabetically.

13.3 Upregulated genes, hypomethylated promoter in Dnmt1$^{-/chip}$

Symbol	CpG	Methylation Dnmt1$^{+/+}$	Methylation Dnmt1$^{-/chip}$	logFC	logCPM	FDR
Clec11a	12	0.73	0.05	2.24	6.426	4.41×10^{-2}
Rec114	41	0.75	0.10	2.48	4.753	1.40×10^{-2}
Slc39a4	20	0.68	0.07	1.44	4.449	8.91×10^{-3}
Plekhg4	50	0.75	0.17	3.71	5.059	1.16×10^{-4}
Tdrd9	33	0.58	0.06	4.02	6.770	8.32×10^{-3}
Scarf1	16	0.65	0.14	1.92	2.467	4.06×10^{-2}
Mrgpre	16	0.92	0.43	4.28	1.339	1.72×10^{-7}
5830444B04Rik	46	0.91	0.44	3.09	−0.251	7.89×10^{-3}
Gm15910	32	0.57	0.11	2.85	1.469	7.26×10^{-3}
Adgrg1	17	0.59	0.13	4.38	−0.873	2.34×10^{-5}
Nxpe5	4	0.52	0.07	2.38	2.399	4.78×10^{-3}
Platr4	20	0.80	0.36	2.95	0.097	2.06×10^{-2}
Gm8008	41	0.71	0.29	3.89	−0.215	1.09×10^{-4}
Lncppara	22	0.77	0.35	2.34	2.640	8.21×10^{-4}
G730013B05Rik	30	0.50	0.11	2.35	−0.723	6.39×10^{-4}
Unc5cl	32	0.50	0.12	5.05	0.574	1.81×10^{-10}
Xlr4a	30	0.53	0.19	2.14	4.974	1.29×10^{-3}
Gm6329	20	0.83	0.50	6.98	1.390	1.83×10^{-6}
Gm10451	24	0.31	0.00	1.81	3.697	1.16×10^{-3}
Lamb2	24	0.52	0.22	1.47	1.640	3.43×10^{-2}
4921525O09Rik	18	0.63	0.33	4.47	0.289	2.18×10^{-11}
Sycp2	40	0.83	0.54	7.08	4.527	9.80×10^{-9}
Gm8773	24	0.29	0.01	2.26	2.793	1.75×10^{-4}
Gm13544	24	0.78	0.51	3.96	−1.418	1.04×10^{-2}
Il18bp	14	0.46	0.19	1.01	3.412	3.07×10^{-2}
Sh2d4a	18	0.88	0.61	3.26	−1.308	9.12×10^{-3}
Xlr4c	23	0.53	0.27	2.10	4.857	2.61×10^{-5}
Igsf23	21	0.69	0.43	3.73	1.204	5.50×10^{-10}
Tex19.1	47	0.87	0.62	8.24	1.553	5.11×10^{-10}
Ceacam19	13	0.27	0.02	1.62	0.428	4.44×10^{-4}
Il2ra	14	0.30	0.06	4.86	0.381	2.61×10^{-5}
D630039A03Rik	26	0.50	0.26	4.39	−0.791	4.69×10^{-3}
Platr17	9	0.35	0.12	2.00	3.331	1.99×10^{-2}
Rnf17	45	0.70	0.47	4.53	4.680	5.64×10^{-3}

Table S13.3 continued on next page.

Symbol	CpG	Methylation Dnmt1$^{+/+}$	Methylation Dnmt1$^{-/chip}$	logFC	logCPM	FDR
Nos1	29	0.83	0.60	3.09	1.995	3.91×10^{-4}
Abhd11os	20	0.28	0.06	1.55	2.050	5.04×10^{-3}
1700048O20Rik	10	0.65	0.44	2.51	2.402	8.86×10^{-3}
Ctsw	24	0.63	0.42	2.41	0.879	2.13×10^{-3}
Nov	34	0.21	0.00	7.80	5.308	2.89×10^{-6}
Afp	14	0.93	0.72	1.99	2.250	2.34×10^{-2}

Table S13.3: Genes assigned to upregulated transcripts (in RNA-seq, $n = 40$) with hypomethylated promoters. Rows are ordered by degree of methylation loss. Downregulated transcripts with hypomethylated promoter ($n = 9$) are not shown.

13.4 Genes covered by H3K4me3 buffer domain

Symbol	MGI annotation	Buffer domain	logFC	FDR
Actb	Actin, Beta	Common	−0.426	6.17×10^{-2}
Ado	2-aminoethanethiol (cysteamine) Dioxygenase	Only Dnmt +/+	NA	NA
Agpat1	1-acylglycerol-3-phosphate O-acyltransferase 1	Only Dnmt +/+	NA	NA
Alas1	Aminolevulinic Acid Synthase 1	Common	0.841	6.52×10^{-3}
Aldoa	Aldolase A, Fructose-bisphosphate	Only Dnmt +/+	NA	NA
Aprt	Adenine Phosphoribosyl Transferase	Common	NA	NA
Atf4	Activating Transcription Factor 4	Common	NA	NA
Atp5b, Atp5h, Atp5j2, Atp5o	Atp Synthase, Mitochondrial F1 Complex, various subunits	Common	NA	NA
Atp6v0c	Atpase, H+ Transporting, Lysosomal V0 Subunit C	Common	NA	NA
B2m	Beta-2 Microglobulin	Common	0.763	1.42×10^{-1}
Bax	Bcl2-associated X Protein	Only Dnmt -/chip	NA	NA
Bckdha	Branched Chain Ketoacid Dehydrogenase E1, Alpha Polypeptide	Only Dnmt +/+	NA	NA
Bcl10	B Cell Leukemia/lymphoma 10	Only Dnmt +/+	NA	NA
Birc5	Baculoviral Iap Repeat-containing 5	Only Dnmt +/+	NA	NA
Calr	Calreticulin	Common	0.482	1.78×10^{-1}
Cdc34	Cell Division Cycle 34	Only Dnmt +/+	NA	NA
Cdk4	Cyclin-dependent Kinase 4	Common	NA	NA

Table S13.4 continued on next page.

Symbol	MGI annotation	Buffer domain	logFC	FDR
Cdkn1a	Cyclin-dependent Kinase Inhibitor 1a (p21)	Only Dnmt +/+	NA	NA
Cdkn2a	Cyclin-dependent Kinase Inhibitor 2a	Common	NA	NA
Cdkn2c	Cyclin-dependent Kinase Inhibitor 2c	Common	NA	NA
Cebpe	Ccaat/enhancer Binding Protein Epsilon	Common	NA	NA
Cers2	Ceramide Synthase 2	Only Dnmt +/+	NA	NA
Cflar	Casp8 And Fadd-like Apoptosis Regulator	Common	-0.318	1.78×10^{-1}
Chac1	Chac, Cation Transport Regulator 1	Common	NA	NA
Chpf2	Chondroitin Polymerizing Factor 2	Common	NA	NA
Ckb	Creatine Kinase, Brain	Common	NA	NA
Coasy	Coenzyme A Synthase	Common	NA	NA
Cox6a1	Cytochrome c oxidase subunit 6A1	Common	NA	NA
Csnk2b	Casein Kinase 2, Beta Polypeptide	Common	NA	NA
Ctsg	Cathepsin G	Common	NA	NA
Ctsw	Cathepsin W	Only Dnmt -/chip	2.406	2.69×10^{-3}
Cxcl10	Chemokine (c-x-c Motif) Ligand 10	Only Dnmt +/+	NA	NA
Cyc1	Cytochrome C-1	Only Dnmt +/+	NA	NA
Cyp26a1	Cytochrome P450, Family 26, Subfamily A, Polypeptide 1	Common	NA	NA
Dctpp1	Dctp Pyrophosphatase 1	Common	NA	NA
Ddx5	Dead (asp-glu-ala-asp) Box Polypeptide 5	Common	1.385	4.21×10^{-4}
Dgat1	Diacylglycerol O-acyltransferase 1	Only Dnmt +/+	-0.588	1.94×10^{-2}
Dpagt1	Dolichyl-phosphate (udp-n-acetylglucosamine) Acetylglucosaminephosphotransferase 1 (glcnac-1-p Transferase)	Only Dnmt -/chip	NA	NA
Dtymk	Deoxythymidylate Kinase	Common	0.381	7.50×10^{-3}
Dusp6	Dual Specificity Phosphatase 6	Only Dnmt +/+	NA	NA
Egr1	Early Growth Response 1	Only Dnmt +/+	NA	NA
Egr2	Early Growth Response 2	Only Dnmt +/+	NA	NA
Elane	Elastase, Neutrophil Expressed	Common	1.250	4.42×10^{-2}
Eno1	Enolase 1, Alpha Non-neuron	Only Dnmt -/chip	NA	NA
Fam213b	Family With Sequence Similarity 213, Member B	Only Dnmt -/chip	NA	NA
Fos	Fbj Osteosarcoma Oncogene	Common	-1.268	2.19×10^{-2}
Fosb	Fbj Osteosarcoma Oncogene B	Common	NA	NA
Gadd45b	Growth Arrest And Dna-damage-inducible 45 Beta	Common	NA	NA

Table S13.4 continued on next page.

Symbol	MGI annotation	Buffer domain	logFC	FDR
Gadd45g	Growth Arrest And Dna-damage-inducible 45 Gamma	Only Dnmt +/+	−1.079	4.05×10^{-2}
Gapdh	Glyceraldehyde-3-phosphate Dehydrogenase	Common	NA	NA
Gatc	Glutamyl-trna(gln) Amidotransferase, Subunit C	Common	NA	NA
Gmppb	Gdp-mannose Pyrophosphorylase B	Only Dnmt +/+	NA	NA
Gng3	Guanine Nucleotide Binding Protein (g Protein), Gamma 3	Only Dnmt +/+	NA	NA
Gngt2	Guanine Nucleotide Binding Protein (g Protein), Gamma Transducing Activity Polypeptide 2	Only Dnmt +/+	NA	NA
Gpi1	Glucose Phosphate Isomerase 1	Only Dnmt -/chip	NA	NA
Gps2	G Protein Pathway Suppressor 2	Common	NA	NA
Hes1	Hairy And Enhancer Of Split 1 (drosophila)	Common	NA	NA
Hhex	Hematopoietically Expressed Homeobox	Common	NA	NA
Hmbs	Hydroxymethylbilane Synthase	Only Dnmt +/+	NA	NA
Hoxa10	Homeobox A10	Common	NA	NA
Hoxa11	Homeobox A11	Common	−3.778	1.82×10^{-3}
Hoxa9	Homeobox A9	Common	NA	NA
Hyal1	Hyaluronoglucosaminidase 1	Only Dnmt +/+	0.778	7.14×10^{-2}
Hyal2	Hyaluronoglucosaminidase 2	Only Dnmt +/+	NA	NA
Id2	Inhibitor Of Dna Binding 2	Only Dnmt +/+	−1.031	8.25×10^{-2}
Idh3b	Isocitrate Dehydrogenase 3 (nad+) Beta	Only Dnmt -/chip	NA	NA
Ifngr1	Interferon Gamma Receptor 1	Common	NA	NA
Il12a	Interleukin 12a	Common	0.466	7.50×10^{-2}
Il3ra	Interleukin 3 Receptor, Alpha Chain	Common	NA	NA
Impdh2	Inosine 5'-phosphate Dehydrogenase 2	Common	NA	NA
Irf3	Interferon Regulatory Factor 3	Only Dnmt -/chip	NA	NA
Isg15	Isg15 Ubiquitin-like Modifier	Only Dnmt +/+	1.114	6.52×10^{-2}
Isyna1	Myo-inositol 1-phosphate Synthase A1	Common	0.830	3.30×10^{-2}
Lmo2	Lim Domain Only 2	Common	NA	NA
Ltb	Lymphotoxin B	Common	−1.341	1.68×10^{-1}
Lyl1	Lymphoblastomic Leukemia 1	Common	NA	NA
Mat2a	Methionine Adenosyltransferase Ii, Alpha	Only Dnmt -/chip	NA	NA
Mcat	Malonyl Coa:acp Acyltransferase (mitochondrial)	Only Dnmt +/+	NA	NA

Table S13.4 continued on next page.

Symbol	MGI annotation	Buffer domain	logFC	FDR
Mcl1	Myeloid Cell Leukemia Sequence 1	Only Dnmt -/chip	NA	NA
Mgat2	Mannoside Acetylglucosaminyltransferase 2	Only Dnmt +/+	NA	NA
Mmp9	Matrix Metallopeptidase 9	Only Dnmt -/chip	-1.337	1.88×10^{-1}
Mogs	Mannosyl-oligosaccharide Glucosidase	Common	NA	NA
Mpo	Myeloperoxidase	Common	1.289	1.73×10^{-1}
Mrpl10	NA	Only Dnmt +/+	NA	NA
Myd88	Myeloid Differentiation Primary Response Gene 88	Only Dnmt +/+	NA	NA
Naprt	Nicotinate Phosphoribosyltransferase	Only Dnmt +/+	NA	NA
Ndufa6	NADH Dehydrogenase 1 Alpha Subcomplex, 6 (b14)	Common	NA	NA
Ndufb7	NADH Dehydrogenase 1 Beta Subcomplex, 7	Only Dnmt +/+	NA	NA
Ndufs7	NADH Dehydrogenase Fe-s Protein 7	Only Dnmt +/+	NA	NA
Nfkbia	Nuclear Factor Of Kappa Light Polypeptide Gene Enhancer In B Cells Inhibitor, Alpha	Only Dnmt +/+	NA	NA
Nme3	Nme/nm23 Nucleoside Diphosphate Kinase 3	Common	NA	NA
Nxf1	Nuclear Rna Export Factor 1	Only Dnmt +/+	NA	NA
Pgam1	Phosphoglycerate Mutase 1	Only Dnmt +/+	-0.504	3.31×10^{-3}
Pgd	Phosphogluconate Dehydrogenase	Common	-0.537	1.01×10^{-1}
Pgls	6-phosphogluconolactonase	Only Dnmt +/+	NA	NA
Pgp	Phosphoglycolate Phosphatase	Common	NA	NA
Phospho1	Phosphatase, Orphan 1	Only Dnmt +/+	NA	NA
Pidd1	P53 Induced Death Domain Protein 1	Only Dnmt +/+	NA	NA
Polr2l	NA	Common	NA	NA
Ppox	Protoporphyrinogen Oxidase	Only Dnmt +/+	NA	NA
Ppp1ca	Protein Phosphatase 1, Catalytic Subunit, Alpha Isoform	Common	NA	NA
Sgsh	N-sulfoglucosamine Sulfohydrolase (sulfamidase)	Only Dnmt -/chip	NA	NA
Shisa5	Shisa Family Member 5	Only Dnmt +/+	NA	NA
Siva1	Siva1, Apoptosis-inducing Factor	Only Dnmt +/+	NA	NA
Six1	Sine Oculis-related Homeobox 1	Common	NA	NA
Slc29a1	Solute Carrier Family 29 (nucleoside Transporters), Member 1	Only Dnmt +/+	NA	NA

Table S13.4 continued on next page.

Symbol	MGI annotation	Buffer domain	logFC	FDR
Smpd2	Sphingomyelin Phosphodiesterase 2, Neutral	Only Dnmt -/chip	NA	NA
Socs3	Suppressor Of Cytokine Signaling 3	Common	NA	NA
Sphk2	Sphingosine Kinase 2	Common	NA	NA
Spi1	Spleen Focus Forming Virus (sffv) Proviral Integration Oncogene	Only Dnmt +/+	NA	NA
Srf	Serum Response Factor	Only Dnmt +/+	NA	NA
Srm	Spermidine Synthase	Only Dnmt -/chip	NA	NA
Srsf2	Serine/arginine-rich Splicing Factor 2	Only Dnmt -/chip	NA	NA
Srsf5	Serine/arginine-rich Splicing Factor 5	Only Dnmt -/chip	NA	NA
Taf10	Tata-box Binding Protein Associated Factor 10	Only Dnmt +/+	NA	NA
Tap2	Transporter 2, Atp-binding Cassette, Sub-family B (mdr/tap)	Only Dnmt -/chip	NA	NA
Tcirg1	T Cell, Immune Regulator 1, Atpase, H+ Transporting, Lysosomal V0 Protein A3	Only Dnmt +/+	NA	NA
Tm7sf2	Transmembrane 7 Superfamily Member 2	Only Dnmt +/+	NA	NA
Tnfrsf14	Tumor Necrosis Factor Receptor Superfamily, Member 14 (herpesvirus Entry Mediator)	Only Dnmt +/+	NA	NA
Tnfsf14	Tumor Necrosis Factor (ligand) Superfamily, Member 14	Common	NA	NA
Tpi1	Triosephosphate Isomerase 1	Common	NA	NA
Tradd	Tnfrsf1a-associated Via Death Domain	Common	NA	NA
Tsta3	Tissue Specific Transplantation Antigen P35b	Common	0.374	1.33×10^{-1}
Tuba4a	Tubulin, Alpha 4a	Common	NA	NA
Uba52	Ubiquitin A-52 Residue Ribosomal Protein Fusion Product 1	Only Dnmt +/+	NA	NA
Umps	Uridine Monophosphate Synthetase	Common	NA	NA
Uqcr11	Ubiquinol-cytochrome C Reductase, Complex Iii Subunit Xi	Common	NA	NA
Zfp36	Zinc Finger Protein 36	Common	−0.919	1.32×10^{-1}

Table S13.4: Genes covered by a H3K4me3 buffer domain [↪ section 8.4, p.60]. logFC and FDR columns refer to the foldchange in Dnmt1$^{-/chip}$ vs. Dnmt1$^{+/+}$ determined by RNA-seq. All rows with *NA* were unchanged and had a $FDR = 1$ assigned.

Supplementary Chapter 14

Supplementary tables for enhancer chapters

14.1 Clade testing for accumulation of CAGE-enhancers

Clade	Obs. Enh.	Other Enh.	Exp.Enh.	Odds ratio	χ squared	FDR	CAGE enhancers
I 1	90	177	164.01	0.2649	9.87×10^{1}	$<1 \times 10^{-7}$	depleted
I 2	83	290	229.12	0.1196	2.96×10^{2}	$<1 \times 10^{-7}$	depleted
I 3	109	20	79.24	3.6613	3.00×10^{1}	$<1 \times 10^{-7}$	accumulated
I 4	292	52	211.30	4.3256	9.55×10^{1}	$<1 \times 10^{-7}$	accumulated
I 5	151	12	100.12	8.8817	7.16×10^{1}	$<1 \times 10^{-7}$	accumulated
I 6	132	115	151.72	0.6868	7.21	1	
I 7	112	1	69.41	77.3744	7.02×10^{1}	$<1 \times 10^{-7}$	accumulated
I 8	124	96	135.14	0.7897	2.44	1	
I 9	122	0	74.94	∞	7.99×10^{1}	$<1 \times 10^{-7}$	accumulated
II 1	114	248	59.06	2.5996	6.51×10^{1}	$<1 \times 10^{-7}$	accumulated
II 2	58	431	79.77	0.6646	7.58	0.891 051	
II 3	39	368	66.40	0.5188	1.43×10^{1}	0.023 858	
II 4	107	217	52.86	2.7820	7.00×10^{1}	$<1 \times 10^{-7}$	accumulated
II 5	102	332	70.80	1.6660	1.76×10^{1}	0.004 228	
II 6	18	65	13.54	1.4308	1.41	1	
II 7	22	564	95.60	0.1761	7.65×10^{1}	$<1 \times 10^{-7}$	depleted
II 8	43	296	55.30	0.7298	3.25	1	
II 9	97	106	33.12	5.2388	1.52×10^{2}	$<1 \times 10^{-7}$	accumulated
II 10	62	378	71.78	0.8272	1.58	1	
II 11	21	190	34.42	0.5546	6.07	1	
II 12	8	184	31.32	0.2147	2.07×10^{1}	0.000 755	depleted
II 13	56	355	67.05	0.7939	2.18	1	
II 14	8	139	23.98	0.2877	1.23×10^{1}	0.067 195	

Table S14.1 continued on next page.

Clade	Obs.	Other	Exp.	Odds ratio	χ squared	FDR	CAGE enhancers
III 1	49	428	98.42	0.4139	3.37×10^{1}	$< 1 \times 10^{-7}$	depleted
III 2	210	13	46.01	76.7722	7.64×10^{2}	$< 1 \times 10^{-7}$	accumulated
III 3	98	1004	227.39	0.3136	1.16×10^{2}	$< 1 \times 10^{-7}$	depleted
III 4	84	396	99.04	0.8005	2.96	1	
III 5	107	95	41.68	4.6960	1.32×10^{2}	$< 1 \times 10^{-7}$	accumulated
III 6	16	299	65.00	0.1940	4.85×10^{1}	$< 1 \times 10^{-7}$	depleted
III 7	24	571	122.77	0.1427	1.12×10^{2}	$< 1 \times 10^{-7}$	depleted
III 8	31	384	85.63	0.2903	4.68×10^{1}	$< 1 \times 10^{-7}$	depleted
III 9	64	540	124.63	0.4218	4.13×10^{1}	$< 1 \times 10^{-7}$	depleted
III 10	34	315	72.01	0.3963	2.64×10^{1}	$< 1 \times 10^{-7}$	depleted
III 11	22	132	31.78	0.6336	3.51	1	
III 12	109	4	23.32	116.3837	4.01×10^{2}	$< 1 \times 10^{-7}$	accumulated
III 13	239	0	49.32	∞	9.58×10^{2}	$< 1 \times 10^{-7}$	accumulated
IV 1	57	254	31.14	2.6332	2.89×10^{1}	$< 1 \times 10^{-7}$	accumulated
IV 2	17	372	38.95	0.3359	1.77×10^{1}	0.003 926	
IV 3	35	731	76.70	0.2584	5.00×10^{1}	$< 1 \times 10^{-7}$	depleted
IV 4	42	0	4.21	∞	3.78×10^{2}	$< 1 \times 10^{-7}$	accumulated
V 1	9	799	21.06	0.4019	7.04	1	
V 2	2	67	1.80	1.1164	0.00	1	
V 3	6	140	3.80	1.6148	7.86×10^{-1}	1	
V 4	11	321	8.65	1.2921	4.18×10^{-1}	1	
V 5	15	547	14.65	1.0262	0.00	1	
V 6	5	112	3.05	1.6805	7.17×10^{-1}	1	
V 7	28	481	13.26	2.3051	1.65×10^{1}	0.007 550	
V 8	6	427	11.28	0.5148	2.17	1	
V 9	31	1100	29.47	1.0598	4.09×10^{-2}	1	
V 10	10	659	17.43	0.5512	3.02	1	
V 11	16	809	21.50	0.7234	1.29	1	
V 12	12	655	17.38	0.6707	1.50	1	
V 13	1	339	8.86	0.1071	6.48	1	
V 14	21	799	21.37	0.9808	0.00	1	
V 15	4	337	8.89	0.4356	2.30	1	
V 16	6	318	8.44	0.6987	4.74×10^{-1}	1	
V 17	6	263	7.01	0.8494	3.91×10^{-2}	1	
V 18	19	473	12.82	1.5374	2.70	1	
V 19	20	486	13.19	1.5790	3.26	1	

Table S14.1 continued on next page.

Clade	Obs.	Other	Exp.	Odds ratio	χ squared	FDR	CAGE enhancers
V 20	26	179	5.34	5.8781	7.96×10^{1}	$<1 \times 10^{-7}$	accumulated
V 21	5	285	7.56	0.6494	5.91×10^{-1}	1	
V 22	7	357	9.49	0.7260	4.42×10^{-1}	1	
V 23	4	189	5.03	0.7879	5.83×10^{-2}	1	
V 24	8	335	8.94	0.8893	2.28×10^{-2}	1	
V 25	4	62	1.72	2.4314	1.90	1	
VI 1	14	406	22.14	0.6082	2.92	1	
VI 2	82	957	54.77	1.6508	1.55×10^{1}	0.012 231	
VI 3	23	586	32.11	0.6904	2.61	1	
VI 4	21	354	19.77	1.0690	2.97×10^{-2}	1	
VI 5	12	362	19.72	0.5853	2.91	1	
VI 6	27	729	39.86	0.6456	4.41	1	
VI 7	25	380	21.35	1.1922	5.13×10^{-1}	1	
VI 8	21	419	23.20	0.8960	1.38×10^{-1}	1	
VI 9	20	277	15.66	1.3103	1.03	1	
VI 10	42	363	21.35	2.1825	2.10×10^{1}	0.000 755	accumulated
VI 11	11	387	20.98	0.4993	4.73	1	
VI 12	8	224	12.23	0.6357	1.23	1	
VI 13	20	477	26.20	0.7427	1.39	1	
VI 14	15	268	14.92	1.0059	0.00	1	
VI 15	34	618	34.37	0.9877	0.00	1	
VI 16	10	268	14.66	0.6635	1.28	1	
VI 17	27	448	25.04	1.0879	9.47×10^{-2}	1	
VI 18	12	323	17.66	0.6590	1.65	1	
VI 19	28	487	27.15	1.0352	5.05×10^{-3}	1	
VI 20	14	234	13.07	1.0773	1.51×10^{-2}	1	
VI 21	1	58	3.11	0.3084	8.86×10^{-1}	1	
VI 22	13	0	0.69	∞	2.15×10^{2}	$<1 \times 10^{-7}$	accumulated
VII 1	16	474	33.61	0.4349	1.02×10^{1}	0.207 021	
VII 2	17	402	28.74	0.5546	5.10	1	
VII 3	34	590	42.80	0.7615	1.94	1	
VII 4	37	278	21.60	1.8936	1.17×10^{1}	0.095 130	
VII 5	15	590	41.49	0.3187	1.96×10^{1}	0.001 510	
VII 6	80	640	49.38	1.8812	2.26×10^{1}	0.000 302	accumulated
VII 7	24	333	24.48	0.9774	0.00	1	
VII 8	2	254	17.56	0.1023	1.45×10^{1}	0.020 838	

Table S14.1 continued on next page.

Clade	Obs.	Other	Exp.	Odds ratio	χ squared	FDR	CAGE enhancers
VII 9	69	408	32.71	2.5808	4.59×10^{1}	$< 1 \times 10^{-7}$	accumulated
VII 10	20	507	36.14	0.5102	8.02	0.696 714	
VII 11	12	461	32.44	0.3327	1.44×10^{1}	0.022 650	
VII 12	38	0	2.61	∞	5.05×10^{2}	$< 1 \times 10^{-7}$	accumulated
VII 13	20	278	20.44	0.9758	0.00	1	
VIII1	21	778	26.21	0.7493	1.14	1	
VIII2	28	556	19.16	1.6447	4.52	1	
VIII3	9	515	17.19	0.4734	4.19	1	
VIII4	4	191	6.40	0.6035	6.16×10^{-1}	1	
VIII5	3	138	4.62	0.6312	2.95×10^{-1}	1	
VIII6	12	283	9.68	1.2801	3.89×10^{-1}	1	
VIII7	3	154	5.15	0.5628	5.73×10^{-1}	1	
VIII8	4	203	6.79	0.5656	8.49×10^{-1}	1	
VIII9	19	272	9.55	2.2739	9.49	0.312 419	
VIII10	7	242	8.17	0.8432	6.08×10^{-2}	1	
VIII11	3	0	0.10	∞	6.07×10^{1}	$< 1 \times 10^{-7}$	accumulated
IX 1	53	0	2.16	∞	1.23×10^{3}	$< 1 \times 10^{-7}$	accumulated
IX 2	16	367	15.63	1.0260	0.00	1	
IX 3	20	526	22.29	0.8851	1.63×10^{-1}	1	
IX 4	19	581	24.49	0.7513	1.16	1	
IX 5	32	875	37.02	0.8408	6.66×10^{-1}	1	
IX 6	45	800	34.49	1.3847	3.46	1	
IX 7	19	562	23.72	0.7792	8.55×10^{-1}	1	
IX 8	1	316	12.94	0.0710	1.11×10^{1}	0.132 880	
IX 9	4	209	8.69	0.4416	2.18	1	
IX 10	4	210	8.74	0.4394	2.21	1	
IX 11	4	288	11.92	0.3164	5.03	1	
IX 12	13	457	19.19	0.6520	1.89	1	
IX 13	28	431	18.74	1.5863	4.59	1	
IX 14	6	323	13.43	0.4239	3.92	1	
IX 15	5	125	5.31	0.9388	0.00	1	
IX 16	5	289	12.00	0.3956	3.84	1	
IX 17	1	103	4.25	0.2253	1.88	1	
X 1	770	266	721.99	1.2962	1.17×10^{1}	0.094 828	
X 2	306	131	304.55	1.0168	1.04×10^{-2}	1	
X 3	293	152	310.12	0.8302	3.10	1	

Table S14.1 continued on next page.

Clade	Obs.	Other	Exp.	Odds ratio	χ squared	FDR	CAGE enhancers
X 4	387	96	336.60	1.8041	2.58×10^{1}	$< 1 \times 10^{-7}$	accumulated
X 5	437	329	533.83	0.5452	6.28×10^{1}	$< 1 \times 10^{-7}$	depleted
X 6	358	148	352.63	1.0553	2.35×10^{-1}	1	
X 7	480	182	461.35	1.1595	2.55	1	
X 8	491	203	483.65	1.0565	3.47×10^{-1}	1	
X 9	459	219	472.50	0.9044	1.28	1	
X 10	261	59	223.01	1.9651	2.16×10^{1}	0.000 453	accumulated
X 11	166	23	131.71	3.1986	2.92×10^{1}	$< 1 \times 10^{-7}$	accumulated
X 12	268	201	326.85	0.5607	3.63×10^{1}	$< 1 \times 10^{-7}$	depleted
X 13	346	2	242.52	79.6872	1.50×10^{2}	$< 1 \times 10^{-7}$	accumulated
X 14	188	9	137.29	9.3410	6.20×10^{1}	$< 1 \times 10^{-7}$	accumulated
X 15	16	110	87.81	0.0608	1.94×10^{2}	$< 1 \times 10^{-7}$	depleted
X 16	215	89	211.86	1.0525	1.13×10^{-1}	1	
X 17	71	13	58.54	2.3915	8.14	0.653 981	
X 18	80	233	218.13	0.1381	2.97×10^{2}	$< 1 \times 10^{-7}$	depleted
X 19	36	20	39.03	0.7816	5.43×10^{-1}	1	
X 20	52	11	43.90	2.0650	4.37	1	
X 21	208	80	200.71	1.1354	7.84×10^{-1}	1	
X 22	85	43	89.20	0.8577	5.15×10^{-1}	1	
X 23	150	44	135.20	1.4948	5.10	1	

Table S14.1: List of enhancer clades including results of the enrichment testing.

14.2 Top 100 enhancer promoter interactions

Enhancer	Transcript	Relevance	Specificity	Clade	Accumulation
chr8:129118460-129118806	Irf2bp2 NM _001164598	123.66	Common	III 12	strong accumulation
chr19:32830462-32830556	Pten NM _008960	98.83	*-/chip*	III 12	strong accumulation
chr5:32437783-32438223	Fosl2 NM _008037	96.25	Common	V 20	strong accumulation
chr2:116948861-116948982	Spred1 NM _033524	66.00	Common	I 1	mild depletion
chr7:31933862-31934002	Gramd1a NM _027898	62.20	Common	I 2	strong depletion
chr5:88982505-88982705	Utp3 NM _023054	57.75	Common	VII 12	strong accumulation
chr6:108609435-108609657	Bhlhe40 NM _011498	55.75	Common	III 5	strong accumulation
chr17:17762016-17762488	Lnpep NM _172827	55.00	*+/+*	I 9	strong accumulation
chr18:75526293-75526427	Smad7 NM _001042660	51.83	Common	III 2	strong accumulation
chr9:45822492-45822694	Sik3 NM _027498	51.50	Common	I 7	strong accumulation
chrX:12859010-12859107	Ddx3x NM _010028	49.11	*+/+*	IX 1	strong accumulation
chr10:20878803-20879105	Myb NM _001198914	43.00	Common	IX 1	strong accumulation
chr11:59919286-59919569	Rai1 NM _009021	42.83	*-/chip*	VII 12	strong accumulation
chr7:148185386-148185532	Ifitm6 NM _001033632	42.00	*+/+*	I 8	n.s.
chrX:48371818-48372117	Rap2c NM _172413	42.00	Common	VII 12	strong accumulation
chr7:109393827-109394241	Rhog NM _019566	41.86	*-/chip*	I 6	n.s.
chr6:97161647-97161930	Arl6ip5 NM _022992	41.00	*+/+*	III 13	strong accumulation
chr9:106270028-106270347	Dusp7 NM _153459	40.83	*-/chip*	III 2	strong accumulation
chr1:173161865-173162317	Fcer1g NM _010185	40.50	*-/chip*	I 2	strong depletion

Table S14.3 continued on next page.

Enhancer	Transcript	Relevance	Specificity	Clade	Accumulation
chr16:8667505-8667642	Carhsp1 NM _025821	40.50	Common	VI 10	mild accumulation
chr12:70781562-70782005	Sos2 NM _001135559	39.50	Common	III 2	strong accumulation
chr3:109144595-109144844	Vav3 NM _020505	39.33	Common	III 12	strong accumulation
chr1:69729203-69729392	Ikzf2 NM _011770	38.88	*+/+*	III 11	n.s.
chr1:69735657-69735883	Ikzf2 NM _011770	38.88	Common	III 11	n.s.
chr4:48056053-48056434	Nr4a3 NM _015743	38.50	*+/+*	III 13	strong accumulation
chr4:48055504-48055879	Nr4a3 NM _015743	38.50	*+/+*	III 13	strong accumulation
chr10:98724349-98724762	Dusp6 NM _026268	37.83	*+/+*	III 12	strong accumulation
chr1:174078988-174079160	Dcaf8 NM _153555	37.00	*-/chip*	III 13	strong accumulation
chr9:40608086-40608407	Hspa8 NM _031165	36.50	Common	III 13	strong accumulation
chr2:165808072-165808206	Ncoa3 NM _008679	36.00	*+/+*	I 7	strong accumulation
chr14:52642893-52643030	Supt16 NM _033618	36.00	*-/chip*	IV 4	strong accumulation
chr6:91060822-91061195	Nup210 NM _018815	36.00	*+/+*	I 7	strong accumulation
chr5:31904815-31905095	Mrpl33 NM _025796	36.00	Common	I 6	n.s.
chr7:109397023-109397354	Rhog NM _019566	35.86	*+/+*	III 2	strong accumulation
chr7:108458928-108459159	Atg16l2 NM _001111111	35.66	*+/+*	I 5	strong accumulation
chr5:106131055-106131485	Lrrc8d NM _178701	35.60	*-/chip*	I 8	n.s.
chr10:41997413-41997753	Foxo3 NM _019740	35.50	Common	III 13	strong accumulation
chr13:102537353-102537458	Pik3r1 NM _001077495	35.33	Common	IX 1	strong accumulation
chr14:55331772-55332002	Cebpe NM _207131	35.00	*+/+*	II 4	mild accumulation
chr12:92826605-92827011	Gtf2a1 NM _031391	34.58	*-/chip*	III 13	strong accumulation

Table S14.3 continued on next page.

Enhancer	Transcript	Relevance	Specificity	Clade	Accumulation
chr16:30597190-30597413	Fam43a NM _177632	33.58	*-/chip*	II 8	n.s.
chr16:4558797-4559125	Tfap4 NM _031182	32.33	*+/+*	III 2	strong accumulation
chr5:53980023-53980438	Rbpj NM _001080927	32.23	*-/chip*	III 2	strong accumulation
chr9:74919909-74920035	Myo5a NM _010864	32.00	*+/+*	III 5	strong accumulation
chr2:174263313-174263398	Ctsz NM _022325	32.00	*-/chip*	III 5	strong accumulation
chr5:97223338-97223805	Anxa3 NM _013470	32.00	*-/chip*	II 1	mild accumulation
chr10:107600249-107600580	Ppp1r12a NM _027892	31.83	Common	III 13	strong accumulation
chr1:166065453-166065906	Sell NM _011346	31.67	*+/+*	I 4	strong accumulation
chr1:166066176-166066373	Sell NM _011346	31.67	Common	I 4	strong accumulation
chr17:35188219-35188359	Clic1 NM _033444	31.48	Common	III 12	strong accumulation
chr11:68910617-68910749	Per1 NM _011065	31.00	Common	I 9	strong accumulation
chr15:34013466-34013685	Mtdh NM _026002	30.75	*-/chip*	III 13	strong accumulation
chr17:8070369-8070499	Tagap NM _145968	30.00	*-/chip*	X 1	n.s.
chr7:109415207-109415438	Rhog NM _019566	30.00	Common	IV 4	strong accumulation
chr11:4132648-4133057	Osm NM _001013365	30.00	*+/+*	II 13	n.s.
chr15:78076647-78076872	Ncf4 NM _008677	30.00	*-/chip*	III 5	strong accumulation
chr8:3355499-3355866	A430078G23Rik NM _001033378	30.00	Common	I 4	strong accumulation
chr7:4452238-4452537	Ppp1r12c NM _029834	29.83	Common	III 13	strong accumulation
chr19:47993077-47993274	Itprip NM _001001738	29.00	Common	III 13	strong accumulation
chr3:51906595-51906750	Maml3 NM _001004176	28.50	*-/chip*	IX 1	strong accumulation
chr8:47827018-47827210	Irf2 NM _008391	28.07	*+/+*	III 2	strong accumulation

Table S14.3 continued on next page.

Enhancer	Transcript	Relevance	Specificity	Clade	Accumulation
chr8:47827018-47827210	Gm16675 NR _045750	28.07	*+/+*	III 2	strong accumulation
chr17:5491287-5491653	Zdhhc14 NM _146073	27.83	*-/chip*	IV 1	mild accumulation
chr16:20098132-20098414	Klhl24 NM _029436	27.50	*-/chip*	III 13	strong accumulation
chr1:186557236-186557391	Hlx NM _008250	27.00	Common	I 6	n.s.
chr11:49840278-49840659	Rnf130 NM _001290750	26.65	*+/+*	I 6	n.s.
chr11:33964195-33964631	Lcp2 NM _010696	26.25	Common	I 3	mild accumulation
chr7:71081166-71081577	Klf13 NM _021366	26.08	*+/+*	III 13	strong accumulation
chr7:71081929-71082352	Klf13 NM _021366	26.08	*-/chip*	IX 1	strong accumulation
chr13:49402928-49403287	Fgd3 NM _015759	25.83	Common	I 4	strong accumulation
chr10:98731320-98731650	Dusp6 NM _026268	25.83	*+/+*	I 2	strong depletion
chrX:53708963-53709226	Ints6l NM _172779	25.50	Common	IX 1	strong accumulation
chr1:133036164-133036533	Dyrk3 NM _145508	25.00	Common	I 5	strong accumulation
chr9:123983194-123983467	Ccr2 NM _009915	25.00	*-/chip*	I 7	strong accumulation
chr2:127161177-127161343	Dusp2 NM _010090	24.50	*-/chip*	III 12	strong accumulation
chr17:84580878-84581346	Zfp36l2 NM _001001806	24.33	*-/chip*	I 9	strong accumulation
chr17:29626025-29626236	Pim1 NM _008842	24.00	Common	I 9	strong accumulation
chr7:133613557-133613992	Lat NM _010689	24.00	Common	I 4	strong accumulation
chr9:107535725-107535907	Ifrd2 NM _025903	24.00	Common	III 2	strong accumulation
chr3:106588127-106588399	Cd53 NM _007651	24.00	*+/+*	I 4	strong accumulation
chr15:50529317-50529724	Trps1 NM _032000	24.00	*-/chip*	I 4	strong accumulation
chr12:114078996-114079309	Pld4 NM _178911	24.00	*+/+*	II 10	n.s.

Table S14.3 continued on next page.

Enhancer	Transcript	Relevance	Specificity	Clade	Accumulation
chr17:50609860-50610013	Plcl2 NM _013880	24.00	Common	I 9	strong accumulation
chrX:93163268-93163451	Msn NM _010833	24.00	*-/chip*	X 5	mild depletion
chr11:103110944-103111083	Map3k14 NM _016896	24.00	*-/chip*	I 8	n.s.
chr12:88229679-88229907	Irf2bpl NM _145836	24.00	*-/chip*	I 4	strong accumulation
chr5:31189576-31189993	Emilin1 NM _133918	24.00	*-/chip*	III 2	strong accumulation
chr6:21952938-21953129	Cped1 NM _001081351	24.00	*+/+*	VI 5	n.s.
chr8:125226628-125226752	Cbfa2t3 NM _009824	24.00	*+/+*	I 2	strong depletion
chr17:24874541-24874763	Msrb1 NM _001346668	23.75	Common	III 2	strong accumulation
chr6:125300570-125301065	Tnfrsf1a NM _011609	23.58	*+/+*	III 5	strong accumulation
chr6:125301422-125301587	Tnfrsf1a NM _011609	23.58	Common	III 5	strong accumulation
chr19:37510308-37510573	Hhex NM _008245	23.50	*-/chip*	III 12	strong accumulation
chr19:53388365-53388572	Mxi1 NM _001008542	23.35	Common	III 13	strong accumulation
chr19:53387512-53387918	Mxi1 NM _001008542	23.35	*-/chip*	III 13	strong accumulation
chr17:66120229-66120539	Rab31 NM _133685	22.83	*-/chip*	I 8	n.s.
chr6:131337475-131337768	Ybx3 NM _139117	22.61	Common	III 2	strong accumulation
chr12:77638450-77638753	Plekhg3 NM _153804	22.00	Common	I 6	n.s.
chr1:137028322-137028434	Lgr6 NM _001033409	22.00	*+/+*	I 4	strong accumulation
chr1:137027723-137027895	Lgr6 NM _001033409	22.00	Common	I 4	strong accumulation

Table S14.3: List of the top 100 of 11534 enhancer promoter interactions established based on HPC-7 Hi-C data. For details regarding the procedure, see section 12.1 on page 95

14.3 Top 100 genes by enhancer enrichment

Gene	Assigned enhancers	Cumulative Confidence Score	Specificity
Irf2bp2	5	144.66	4x Common, 1x +/+
Rhog	4	125.58	2x Common, 1x -/*chip*, 1x +/+
Pten	2	100.16	1x Common, 1x -/*chip*
Fosl2	1	96.25	1x Common
Sell	4	87.50	2x Common, 2x +/+
Mxi1	3	87.19	2x Common, 1x -/*chip*
Clic1	4	81.48	4x Common
Lrrc8d	4	79.97	1x Common, 3x -/*chip*
Ikzf2	2	77.76	1x Common, 1x +/+
Nr4a3	2	77.00	2x +/+
Gramd1a	2	68.20	1x Common, 1x -/*chip*
Gpank1	5	68.13	5x Common
Bex6	4	66.00	4x Common
Spred1	1	66.00	1x Common
Dusp6	2	63.66	2x +/+
Smpdl3b	4	63.00	1x Common, 2x -/*chip*, 1x +/+
Zeb2	2	62.15	2x -/*chip*
Ncf4	3	60.00	2x -/*chip*, 1x +/+
Six1	3	60.00	1x Common, 2x -/*chip*
Utp3	1	57.75	1x Common
Map3k1	4	57.00	2x Common, 1x -/*chip*, 1x +/+
Bhlhe40	1	55.75	1x Common
Lnpep	1	55.00	1x +/+
Yy1	4	55.00	3x Common, 1x +/+
Spi1	3	54.60	2x Common, 1x -/*chip*
Smad7	3	53.83	2x Common, 1x -/*chip*
Jade2	3	52.23	2x Common, 1x +/+
Klf13	2	52.16	1x -/*chip*, 1x +/+
Sik3	1	51.50	1x Common
Elmo1	4	50.81	1x Common, 1x -/*chip*, 2x +/+
Zfp36l2	5	49.66	2x Common, 2x -/*chip*, 1x +/+
Ddx3x	1	49.11	1x +/+
Tagap	2	48.00	2x -/*chip*

Table S14.4 continued on next page.

Gene	Assigned enhancers	Cumulative Confidence Score	Specificity
Tnfrsf1a	2	47.16	1x Common, 1x *+/+*
Cebpe	2	47.00	1x Common, 1x *+/+*
Itprip	2	47.00	2x Common
Lmo2	2	47.00	1x Common, 1x *+/+*
Larp1	3	46.00	2x Common, 1x *+/+*
Lgr6	2	44.00	1x Common, 1x *+/+*
Ccr2	5	43.00	3x Common, 2x *-/chip*
Myb	1	43.00	1x Common
Rnf19b	6	43.00	1x Common, 4x *-/chip*, 1x *+/+*
Rai1	1	42.83	1x *-/chip*
Tbl1xr1	4	42.47	1x *-/chip*, 3x *+/+*
Dnajc1	3	42.00	2x Common, 1x *+/+*
Ifitm6	1	42.00	1x *+/+*
Ncoa3	2	42.00	1x *-/chip*, 1x *+/+*
Rap2c	1	42.00	1x Common
Dusp7	2	41.83	2x *-/chip*
Nup98	4	41.79	2x Common, 1x *-/chip*, 1x *+/+*
Gon7	3	41.68	1x Common, 2x *-/chip*
Ubr7	3	41.68	1x Common, 2x *-/chip*
Osm	2	41.58	1x Common, 1x *+/+*
Arl6ip5	1	41.00	1x *+/+*
Egr1	3	41.00	2x Common, 1x *-/chip*
L3mbtl3	2	41.00	1x Common, 1x *-/chip*
Carhsp1	1	40.50	1x Common
Fcer1g	1	40.50	1x *-/chip*
Gm16675	2	40.39	1x Common, 1x *+/+*
Irf2	2	40.39	1x Common, 1x *+/+*
Slc38a1	2	39.99	2x Common
Sos2	1	39.50	1x Common
Vav3	1	39.33	1x Common
Per1	2	37.42	1x Common, 1x *-/chip*
Dcaf8	1	37.00	1x *-/chip*
Msn	2	37.00	2x *-/chip*
Atg16l2	2	36.66	1x Common, 1x *+/+*
Pik3r1	2	36.66	1x Common, 1x *+/+*
Hspa8	1	36.50	1x Common

Table S14.4 continued on next page.

Gene	Assigned enhancers	Cumulative Confidence Score	Specificity
Mcl1	3	36.00	3x Common
Mrpl33	1	36.00	1x Common
Nup210	1	36.00	1x *+/+*
Pim1	2	36.00	1x Common, 1x *-/chip*
Pstpip1	2	36.00	2x *-/chip*
Supt16	1	36.00	1x *-/chip*
Foxo3	1	35.50	1x Common
Gtf2a1	1	34.58	1x *-/chip*
Tsc22d3	3	34.50	2x Common, 1x *-/chip*
Fam43a	1	33.58	1x *-/chip*
Tfap4	2	33.33	1x Common, 1x *+/+*
Srgn	4	33.00	3x Common, 1x *-/chip*
Gnptab	4	32.99	2x Common, 2x *-/chip*
Parp8	5	32.50	2x Common, 2x *-/chip*, 1x *+/+*
Rbpj	1	32.23	1x *-/chip*
Anxa3	1	32.00	1x *-/chip*
Ctsz	1	32.00	1x *-/chip*
Myo5a	1	32.00	1x *+/+*
Ppp1r12a	1	31.83	1x Common
Fbxo45	4	31.02	4x Common
Wdr53	4	31.02	4x Common
Mtdh	1	30.75	1x *-/chip*
Dusp2	2	30.50	2x *-/chip*
Plac8	2	30.33	1x Common, 1x *+/+*
5430416N02Rik	3	30.00	1x Common, 1x *-/chip*, 1x *+/+*
A430078G23Rik	1	30.00	1x Common
Ccr1	2	30.00	2x Common
Gse1	1	30.00	1x Common
Hhex	2	30.00	2x *-/chip*
Pmaip1	5	30.00	2x Common, 3x *+/+*
Samsn1	3	30.00	3x Common

Table S14.4: Tabular representation of the top 100 enhancer-targeted genes ranked by cumulative confidence score of all enhancers and transcripts involved.

14.4 Cloned sgRNAs for CRISPRi experiments

ID	Source	PAM sequence	Target
ZsgS _00000	Weissman _sgRNAlist _V4 sgGFP-NT2	GACCAGGATGGGCACCACCC	GFP / non-targeting control
ZsgS _00001	Weissman _mCRISPRiv2 non-targeting _00000	GGGAACCACATGGAATTCGA	non-targeting control
ZsgS _00002	Weissman _mCRISPRiv2 non-targeting _00001	GAGGTTACCCACCCAGCGGT	non-targeting control
ZsgS _00003	designed with CCTop [398]	TGGCGGCGGCCTTGCGTTAG	Intergenic Enhancer Hoxa Locus chr6:52178835-52179286
ZsgS _00004	designed with CCTop [398]	CAGAGCCTGCCTAAATTCCG	Intronic Enhancer Six1 chr12:74145787-74145937
ZsgS _00005	designed with CCTop [398]	AGGACGACTTCGGCCGGCAA	Intronic Enhancer Evi1/Mecom chr3:30412977-30413205
ZsgS _00006	designed with CCTop [398]	CTGACTCACCGGGCTTGTAC	Intergenic Enhancer Meis1 chr11:18778949-18779138
ZsgS _00007	Weissman _mCRISPRiv2 Zeb1 _- _5591915.23-P1	GACAAGCGAGAGGATCATGG	Promoter Zeb1
ZsgS _00008	Weissman _mCRISPRiv2 Hoxa9 _- _52225840.23-P1P2	GCGCCTGGGGATGCACGTAG	Promoter Hoxa9
ZsgS _00009	Weissman _mCRISPRiv2 Hoxa10 _+ _52240448.23-P1	GTCTGATACTAAGAGCAGCA	Promoter Hoxa10
ZsgS _00010	Weissman _mCRISPRiv2 Meis1 _+ _19018881.23-P1P2	GAAGCGGGCAGCATCGATCG	Promoter Meis1
ZsgS _00011	Weissman _mCRISPRiv2 Mecom _+ _30509452.23-P1P2	GGGCACAGCATGAGATCCAA	Promoter Evi1/Mecom
ZsgS _00012	designed with CCTop [398]	TATTGCCCCCTCGGCAGCGT	Intergenic Enhancer Hoxa Locus chr6:52171588-52172019
ZsgS _00013	designed with CCTop [398]	TGCACTGCACCGGGCCAACT	Intronic Enhancer Pias4 chr10:80632144-80632354
ZsgS _00014	designed with CCTop [398]	AGACGTAATTCGCACGCACA	Intronic Enhancer Ftx-Locus chrX:100811607-100811808

Table S14.5: Details of the sgRNA constructs cloned for the preliminary CRISPRi experiments. [↪ subsection 12.2.3, p.108]

14.5 Transcripts responding to Mll2 deletion mediated putatively by enhancer(s)

Transcript	Gene name	logFC	logRPKM	FDR	Mll2-Enhancer(s)
Transcripts targeted by potential Mll2-binding enhancer(s)					
NM _199299	Jade2	−1.040	3.412	8.70×10^{-23}	chr11:51669596-51669969 chr11:51671485-51671764
NM _009127	Scd1	−1.248	1.959	5.18×10^{-17}	chr19:44462997-44463336
NM _172647	F11r	−4.278	−0.236	1.48×10^{-15}	chr1:173408280-173408407
NM _001004176	Maml3	−2.310	−0.407	2.77×10^{-15}	chr3:51456706-51457015
NM _130450	Elovl6	−0.808	3.623	5.89×10^{-14}	chr3:129236606-129236700
NR _027827		−2.327	0.146	2.45×10^{-7}	chr11:69394935-69395324
NM _009778	C3	−0.396	7.489	3.85×10^{-6}	chr17:57430250-57430630
NM _001029890	Mex3a	−0.827	1.065	3.76×10^{-5}	chr3:88335305-88335468 chr3:88335499-88335889
NM _134188	Acot2	−0.981	2.248	1.04×10^{-4}	chr12:85516606-85516919
NM _008883	Plxna3	−1.202	−0.349	1.38×10^{-4}	chrX:71491058-71491174
NM _010590	Ajuba	−1.309	0.645	2.33×10^{-4}	chr14:55222205-55222447
NM _001163567	Fam102b	−0.416	3.632	3.12×10^{-4}	chr3:109144595-109144844
NM _010688	Lasp1	−0.306	6.374	5.81×10^{-4}	chr11:97661550-97661658
NM _011577	Tgfb1	−0.355	7.666	6.31×10^{-4}	chr7:26474146-26474571
NM _001085390	Dusp5	−1.934	0.346	8.89×10^{-4}	chr19:53495930-53496128
NM _029011	Pyroxd2	0.401	4.688	2.23×10^{-3}	chr19:42827767-42827942
NR _033147		0.849	1.536	3.70×10^{-3}	chr5:139848110-139848361
NM _019566	Rhog	−0.297	7.991	3.76×10^{-3}	chr7:109393827-109394241
NM _007695	Chil1	−1.254	1.248	3.85×10^{-3}	chr1:135995776-135996050
NM _029418	9130401M01Rik	0.371	5.085	3.92×10^{-3}	chr15:57807247-57807408
NM _021462	Mknk2	−0.294	6.754	4.16×10^{-3}	chr10:80599712-80599977 chr10:80032504-80032725 chr10:80140860-80141261
NM _009008	Rac2	−0.303	8.572	4.49×10^{-3}	chr15:78428205-78428571
NM _172685	Slc25a24	−0.377	3.970	5.53×10^{-3}	chr3:109144595-109144844
NM _026058	Cers4	−1.488	−0.289	7.50×10^{-3}	chr8:4677329-4677516 chr8:4677726-4677994
NM _021557	Rdh11	0.278	6.044	1.09×10^{-2}	chr12:81535760-81536235
NM _025388	Ufc1	0.265	7.093	1.28×10^{-2}	chr1:172974741-172974926 chr1:173408280-173408407
NM _175515	Intu	0.451	1.544	1.39×10^{-2}	chr3:40604409-40604530 chr3:40753330-40753766

Table S14.6 continued on next page.

Transcript	Gene name	logFC	logRPKM	FDR	Mll2-Enhancer(s)
NM _178744	Zbtb1	0.251	5.214	1.41×10^{-2}	chr12:77471939-77472123
NM _183390	Klhl6	−0.362	4.026	1.56×10^{-2}	chr16:20098132-20098414
NM _199449	Zhx2	−0.339	3.163	1.63×10^{-2}	chr15:57807247-57807408 chr15:57526896-57527250
NM _033523	Spred2	−0.310	4.392	2.35×10^{-2}	chr11:19820257-19820652
NM _019680	Elf4	−0.228	5.757	2.91×10^{-2}	chrX:45786774-45787018
NM _010500	Ier5	0.316	3.387	3.23×10^{-2}	chr1:156906671-156907073
NM _146019	Chd3	−0.225	5.021	3.54×10^{-2}	chr11:68910617-68910749
Transcripts potentially binding Mll2 at the promoter and enhancer(s)					
NM _012006	Acot1	−1.827	2.088	2.60×10^{-11}	chr12:85516606-85516919
NM _025877	Slc25a23	−2.019	0.868	2.44×10^{-9}	chr17:57430250-57430630
NM _010511	Ifngr1	0.258	8.328	2.73×10^{-3}	chr10:19304051-19304238
NR _045750		−1.194	0.685	1.98×10^{-2}	chr8:47826511-47826652
NM _025983	Atp5e	0.244	8.560	3.17×10^{-2}	chr2:174153877-174154243

Table S14.6: Differentially expressed genes after Mll2 deletion in MLL-AF9 leukemia, which were assigned to an enhancer comprising a putative Mll2-binding motif. Rows are ordered by false discovery rate (FDR).

Bibliography

[1] Cooper GM, Hausman RE. The Complexity of Eukaryotic Genomes. In: The Cell - A Molecular Approach. 2nd ed. Sinauer Associates; 2000. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9846/.

[2] Makaowski W. Genomic scrap yard: how genomes utilize all that junk. Gene. 2000 Dec;259:61–67.

[3] Palazzo AF, Gregory TR. The case for junk DNA. PLoS genetics. 2014;10(5):e1004351.

[4] Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, Pérez-Torres CA, Carretero-Paulet L, Chang TH, et al. Architecture and evolution of a minute plant genome. Nature. 2013 Jun;498:94–98.

[5] Carretero-Paulet L, Librado P, Chang TH, Ibarra-Laclette E, Herrera-Estrella L, Rozas J, et al. High Gene Family Turnover Rates and Gene Space Adaptation in the Compact Genome of the Carnivorous Plant Utricularia gibba. Molecular biology and evolution. 2015 May;32:1284–1295.

[6] Dufresne F, Jeffery N. A guided tour of large genome size in animals: what we know and where we are heading. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology. 2011 Oct;19:925–938.

[7] Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. On the immortality of television sets:function in the human genome according to the evolution-free gospel of ENCODE. Genome biology and evolution. 2013;5(3):578–590.

[8] Doolittle WF. Is junk DNA bunk? A critique of ENCODE. Proceedings of the National Academy of Sciences. 2013;110(14):5294–5300.

[9] Niu DK, Jiang L. Can ENCODE tell us how much junk DNA we carry in our genome? Biochemical and biophysical research communications. 2013 Jan;430:1340–1343.

[10] Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, et al. The transcriptional landscape of the mammalian genome. Science (New York, NY). 2005 Sep;309:1559–1563.

[11] Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012 Sep;489:101–108.

[12] Palazzo AF, Lee ES. Non-coding RNA: what is functional and what is junk? Frontiers in genetics. 2015;6:2.

[13] Rands CM, Meader S, Ponting CP, Lunter G. 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. PLoS genetics. 2014;10(7):e1004525.

[14] Ward LD, Kellis M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science (New York, NY). 2012 Sep;337:1675–1678.

[15] Noordermeer D, Duboule D. Chromatin looping and organization at developmentally regulated gene loci. Wiley interdisciplinary reviews Developmental biology. 2013;2:615–630.

[16] Grubert F, Zaugg JB, Kasowski M, Ursu O, Spacek DV, Martin AR, et al. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. Cell. 2015 Aug;162:1051–1065.

[17] Bertolino E, Reinitz J, Manu. The analysis of novel distal Cebpa enhancers and silencers using a transcriptional model reveals the complex regulatory logic of hematopoietic lineage specification. Developmental biology. 2016 May;413:128–144.

[18] Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell. 2016 Nov;167:1369–1384.e19.

[19] Allahyar A, Vermeulen C, Bouwman BAM, Krijger PHL, Verstegen MJAM, Geeven G, et al. Enhancer hubs and loop collisions identified from single-allele topologies. Nature genetics. 2018 Jul;.

[20] Stampfel G, Kazmar T, Frank O, Wienerroither S, Reiter F, Stark A. Transcriptional regulators form diverse groups with context-dependent regulatory functions. Nature. 2015 Dec;528:147–151.

[21] Marsman J, Horsfield JA. Long distance relationships: enhancer-promoter communication and dynamic gene transcription. Biochim Biophys Acta. 2012;1819(11-12):1217–1227. Available from: http://dx.doi.org/10.1016/j.bbagrm.2012.10.008.

[22] Gibcus JH, Dekker J. The hierarchy of the 3D genome. Mol Cell. 2013 Mar;49(5):773–782. Available from: http://dx.doi.org/10.1016/j.molcel.2013.02.011.

[23] Witzgall R, O'Leary E, Leaf A, Onaldi D, Bonventre JV. The Krüppel-associated box-A (KRAB-A) domain of zinc finger proteins mediates transcriptional repression. Proceedings of the National Academy of Sciences of the United States of America. 1994 May;91:4514–4518.

[24] Urrutia R. KRAB-containing zinc-finger repressor proteins. Genome biology. 2003;4:231.

[25] Ogbourne S, Antalis TM. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. The Biochemical journal. 1998 Apr;331 (Pt 1):1–14.

[26] Frenkel B, Montecino M, Stein JL, Lian JB, Stein GS. A composite intragenic silencer domain exhibits negative and positive transcriptional control of the bone-specific osteocalcin gene: promoter and cell type requirements. Proceedings of the National Academy of Sciences of the United States of America. 1994 Nov;91:10923–10927.

[27] Li YP, Chen W, Stashenko P. Characterization of a silencer element in the first exon of the human osteocalcin gene. Nucleic acids research. 1995 Dec;23:5064–5072.

[28] Kassis JA, Brown JL. Polycomb group response elements in Drosophila and vertebrates. Advances in genetics. 2013;81:83–118.

[29] Bauer M, Trupke J, Ringrose L. The quest for mammalian Polycomb response elements: are we there yet? Chromosoma. 2016 Jun;125:471–496.

[30] Rickels R, Herz HM, Sze CC, Cao K, Morgan MA, Collings CK, et al. Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. Nature genetics. 2017;49(11):1647.

[31] Donze D, Kamakaka RT. Braking the silence: how heterochromatic gene repression is stopped in its tracks. BioEssays : news and reviews in molecular, cellular and developmental biology. 2002 Apr;24:344–349.

[32] Gaszner M, Felsenfeld G. Insulators: exploiting transcriptional and epigenetic mechanisms. Nature reviews Genetics. 2006 Sep;7:703–713.

[33] Barkess G, West AG. Chromatin insulator elements: establishing barriers to set heterochromatin boundaries. Epigenomics. 2012 Feb;4:67–80.

[34] Ong CT, Corces VG. CTCF: an architectural protein bridging genome topology and function. Nature reviews Genetics. 2014;15(4):234.

[35] Walters MC, Fiering S, Bouhassira EE, Scalzo D, Goeke S, Magis W, et al. The chicken beta-globin 5'HS4 boundary element blocks enhancer-mediated suppression of silencing. Molecular and cellular biology. 1999 May;19:3714–3726.

[36] Yao S, Osborne CS, Bharadwaj RR, Pasceri P, Sukonnik T, Pannell D, et al. Retrovirus silencer blocking by the cHS4 insulator is CTCF independent. Nucleic acids research. 2003 Sep;31:5317–5323.

[37] Rincón-Arano H, Furlan-Magaril M, Recillas-Targa F. Protection against telomeric position effects by the chicken cHS4 beta-globin insulator. Proceedings of the National Academy of Sciences of the United States of America. 2007 Aug;104:14044–14049.

[38] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (New York, NY). 2009 Oct;326:289–293.

[39] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012 Apr;485:376–380.

[40] Jost D, Carrivain P, Cavalli G, Vaillant C. Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. Nucleic acids research. 2014 Sep;42:9553–9561.

[41] Sexton T, Cavalli G. The role of chromosome domains in shaping the functional genome. Cell. 2015;160(6):1049–1059.

[42] Hug CB, Grimaldi AG, Kruse K, Vaquerizas JM. Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription. Cell. 2017 Apr;169:216–228.e19.

[43] Flyamer IM, Gassler J, Imakaev M, Brandão HB, Ulianov SV, Abdennur N, et al. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Nature. 2017 Apr;544:110–114.

[44] Nora EP, Goloborodko A, Valton AL, Gibcus JH, Uebersohn A, Abdennur N, et al. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell. 2017 May;169:930–944.e22.

[45] Li Q, Peterson KR, Fang X, Stamatoyannopoulos G. Locus control regions. Blood. 2002 Nov;100:3077–3086.

[46] Bulger M, Groudine M. Locus Control Regions (LCRs). eLS. 2013;Available from: http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0005034.pub2/full.

[47] Forrester WC, Epner E, Driscoll MC, Enver T, Brice M, Papayannopoulou T, et al. A deletion of the human beta-globin locus activation region causes a major alteration in chromatin structure and replication across the entire beta-globin locus. Genes & development. 1990;4(10):1637–1649.

[48] Jiménez G, Griffiths SD, Ford AM, Greaves MF, Enver T. Activation of the beta-globin locus control region precedes commitment to the erythroid lineage. Proceedings of the National Academy of Sciences of the United States of America. 1992 Nov;89:10618–10622.

[49] Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, et al. Super-enhancers in the control of cell identity and disease. Cell. 2013 Nov;155(4):934–947. Available from: http://dx.doi.org/10.1016/j.cell.2013.09.053.

[50] Iacono M, Mignone F, Pesole G. uAUG and uORFs in human and rodent 5 untranslated mRNAs. Gene. 2005;349:97–105.

[51] Kochetov AV, Ahmad S, Ivanisenko V, Volkova OA, Kolchanov NA, Sarai A. uORFs, reinitiation and alternative translation start sites in human mRNAs. FEBS letters. 2008;582(9):1293–1297.

[52] Wang Z, Xiao X, Van Nostrand E, Burge CB. General and specific functions of exonic splicing silencers in splicing control. Molecular cell. 2006 Jul;23:61–70.

[53] Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489(7414):109.

[54] Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nature genetics. 2015 Jun;47:598–606.

[55] Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, et al. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. Science (New York, NY). 2016 Nov;354:769–773.

[56] Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, Oostra BA, et al. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Human molecular genetics. 2003 Jul;12:1725–1735.

[57] Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 2013 Nov;503:290–294.

[58] van Arensbergen J, van Steensel B, Bussemaker HJ. In search of the determinants of enhancer–promoter interaction specificity. Trends in cell biology. 2014;24(11):695–702.

[59] Drissen R, Palstra RJ, Gillemans N, Splinter E, Grosveld F, Philipsen S, et al. The active spatial organization of the beta-globin locus requires the transcription factor EKLF. Genes & development. 2004 Oct;18:2485–2490.

[60] Vakoc CR, Letting DL, Gheldof N, Sawado T, Bender MA, Groudine M, et al. Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. Molecular cell. 2005 Feb;17:453–462.

[61] Hughes JR, Roberts N, McGowan S, Hay D, Giannoulatou E, Lynch M, et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nature genetics. 2014;46(2):205.

[62] Ho Y, Cooke NE, Liebhaber SA. An autoregulatory pathway establishes the definitive chromatin conformation at the pit-1 locus. Mol Cell Biol. 2015 May;35(9):1523–1532. Available from: http://dx.doi.org/10.1128/MCB.01283-14.

[63] Kieffer-Kwon KR, Tang Z, Mathe E, Qian J, Sung MH, Li G, et al. Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. Cell. 2013 Dec;155(7):1507–1520. Available from: http://dx.doi.org/10.1016/j.cell.2013.11.039.

[64] Kaiser K, Meisterernst M. The human general co-factors. Trends in biochemical sciences. 1996 Sep;21:342–345.

[65] Ogryzko VV, Schiltz RL, Russanova V, Howard BH, Nakatani Y. The transcriptional coactivators p300 and CBP are histone acetyltransferases. Cell. 1996 Nov;87:953–959.

[66] Spencer TE, Jenster G, Burcin MM, Allis CD, Zhou J, Mizzen CA, et al. Steroid receptor coactivator-1 is a histone acetyltransferase. Nature. 1997 Sep;389:194–198.

[67] Hirose Y, Ohkuma Y. Phosphorylation of the C-terminal domain of RNA polymerase II plays central roles in the integrated events of eucaryotic gene expression. Journal of biochemistry. 2007 May;141:601–608.

[68] Mittler G, Kremmer E, Timmers HT, Meisterernst M. Novel critical role of a human Mediator complex for basal RNA polymerase II transcription. EMBO reports. 2001 Sep;2:808–813.

[69] Kornberg RD. Mediator and the mechanism of transcriptional activation. Trends in biochemical sciences. 2005 May;30:235–239.

[70] Robinson PJ, Trnka MJ, Bushnell DA, Davis RE, Mattei PJ, Burlingame AL, et al. Structure of a Complete Mediator-RNA Polymerase II Pre-Initiation Complex. Cell. 2016 Sep;166:1411–1422.e16.

[71] C Quaresma AJ, Bugai A, Barboric M. Cracking the control of RNA polymerase II elongation by 7SK snRNP and P-TEFb. Nucleic acids research. 2016 Sep;44:7527–7539.

[72] Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science (New York, NY). 2008 Dec;322:1845–1848.

[73] Jonkers I, Kwak H, Lis JT. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. eLife. 2014 Apr;3:e02407. Original DateCompleted: 20140520.

[74] Flynn RA, Do BT, Rubin AJ, Calo E, Lee B, Kuchelmeister H, et al. 7SK-BAF axis controls pervasive transcription at enhancers. Nature structural & molecular biology. 2016 Mar;23:231–238.

[75] Liu W, Ma Q, Wong K, Li W, Ohgi K, Zhang J, et al. Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. Cell. 2013 Dec;155:1581–1595.

[76] Ong CT, Corces VG. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nature reviews Genetics. 2011 Apr;12:283–293.

[77] Schaukowitch K, Joo JY, Liu X, Watts JK, Martinez C, Kim TK. Enhancer RNA facilitates NELF release from immediate early genes. Mol Cell. 2014 Oct;56(1):29–42. Available from: http://dx.doi.org/10.1016/j.molcel.2014.08.023.

[78] Grosschedl R, Birnstiel M. Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants in vivo. Proceedings of the National Academy of Sciences. 1980;77(3):1432–1436.

[79] Palla F, Bonura C, Anello L, Di Gaetano L, Spinelli G. Modulator factor-binding sequence of the sea urchin early histone H2A promoter acts as an enhancer element. Proceedings of the National Academy of Sciences of the United States of America. 1994 Dec;91:12322–12326.

[80] Palla F, Melfi R, Anello L, Di Bernardo M, Spinelli G. Enhancer blocking activity located near the 3′ end of the sea urchin early H2A histone gene. Proceedings of the National Academy of Sciences of the United States of America. 1997 Mar;94:2272–2277.

[81] West RW, Yocum RR, Ptashne M. Saccharomyces cerevisiae GAL1-GAL10 divergent promoter region: location and function of the upstream activating sequence UASG. Molecular and cellular biology. 1984 Nov;4:2467–2478.

[82] Bram RJ, Lue NF, Kornberg RD. A GAL family of upstream activating sequences in yeast: roles in both induction and repression of transcription. The EMBO journal. 1986 Mar;5:603–608.

[83] Webster N, Jin JR, Green S, Hollis M, Chambon P. The yeast UASG is a transcriptional enhancer in human HeLa cells in the presence of the GAL4 trans-activator. Cell. 1988 Jan;52:169–178.

[84] Morett E, Segovia L. The sigma 54 bacterial enhancer-binding protein family: mechanism of action and phylogenetic relationship of their functional domains. Journal of bacteriology. 1993 Oct;175:6067–6074.

[85] Khoury G, Gruss P. Enhancer elements. Cell. 1983;33(2):313–314.

[86] Zenke M, Grundström T, Matthes H, Wintzerith M, Schatz C, Wildeman A, et al. Multiple sequence motifs are involved in SV40 enhancer function. The EMBO journal. 1986 Feb;5:387–397.

[87] Ondek B, Gloss L, Herr W. The SV40 enhancer contains two distinct levels of organization. Nature. 1988 May;333:40–45.

[88] Banerji J, Olson L, Schaffner W. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. Cell. 1983 Jul;33:729–740.

[89] Gillies SD, Morrison SL, Oi VT, Tonegawa S. A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. Cell. 1983;33(3):717–728.

[90] Grosveld F, van Assendelft GB, Greaves DR, Kollias G. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. Cell. 1987 Dec;51:975–985.

[91] Greaves DR, Wilson FD, Lang G, Kioussis D. Human CD2 3′-flanking sequences confer high-level, T cell-specific, position-independent gene expression in transgenic mice. Cell. 1989 Mar;56:979–986.

[92] Batut PJ, Gingeras TR. Conserved noncoding transcription and core promoter regulatory code in early , javax.xml.bind.JAXBElement@11bcf85, development. eLife. 2017 Dec;6.

[93] Heinz S, Romanoski CE, Benner C, Glass CK. The selection and function of cell type-specific enhancers. Nat Rev Mol Cell Biol. 2015 Feb;Available from: http://dx.doi.org/10.1038/nrm3949.

[94] Mayran A, Khetchoumian K, Hariri F, Pastinen T, Gauthier Y, Balsalobre A, et al. Pioneer factor Pax7 deploys a stable enhancer repertoire for specification of cell fate. Nature genetics. 2018 Feb;50:259–269.

[95] Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A. 2010 Dec;107(50):21931–21936. Available from: http://dx.doi.org/10.1073/pnas.1016071107.

[96] Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 2010 Aug;28(8):817–825. Available from: http://dx.doi.org/10.1038/nbt.1662.

[97] Calo E, Wysocka J. Modification of enhancer chromatin: what, how, and why? Mol Cell. 2013 Mar;49(5):825–837. Available from: http://dx.doi.org/10.1016/j.molcel.2013.01.038.

[98] Collis P, Antoniou M, Grosveld F. Definition of the minimal requirements within the human beta-globin gene and the dominant control region for high level expression. The EMBO journal. 1990 Jan;9:233–240.

[99] Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010 May;465(7295):182–187. Available from: http://dx.doi.org/10.1038/nature09033.

[100] De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, et al. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol. 2010 May;8(5):e1000384. Available from: http://dx.doi.org/10.1371/journal.pbio.1000384.

[101] Zhu Y, Sun L, Chen Z, Whitaker JW, Wang T, Wang W. Predicting enhancer transcription and activity from chromatin modifications. Nucleic Acids Res. 2013 Dec;41(22):10032–10043. Available from: http://dx.doi.org/10.1093/nar/gkt826.

[102] Koch F, Fenouil R, Gut M, Cauchy P, Albert TK, Zacarias-Cabeza J, et al. Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. Nature structural & molecular biology. 2011 Jul;18:956–963.

[103] Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014 Mar;507(7493):455–461. Available from: http://dx.doi.org/10.1038/nature12787.

[104] Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature genetics. 2014 Dec;46:1311–1320.

[105] Wang D, Garcia-Bassets I, Benner C, Li W, Su X, Zhou Y, et al. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature. 2011 May;474:390–394.

[106] Li W, Notani D, Ma Q, Tanasa B, Nunez E, Chen AY, et al. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature. 2013 Jun;498(7455):516–520. Available from: http://dx.doi.org/10.1038/nature12210.

[107] Kaikkonen MU, Spann NJ, Heinz S, Romanoski CE, Allison KA, Stender JD, et al. Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription. Molecular cell. 2013 Aug;51:310–325.

[108] Aguilo F, Li S, Balasubramaniyan N, Sancho A, Benko S, Zhang F, et al. Deposition of 5-Methylcytosine on Enhancer RNAs Enables the Coactivator Function of PGC-1α. Cell reports. 2016 Jan;14:479–492.

[109] Lam MTY, Li W, Rosenfeld MG, Glass CK. Enhancer RNAs and regulated transcriptional programs. Trends Biochem Sci. 2014 Apr;39(4):170–182. Available from: http://dx.doi.org/10.1016/j.tibs.2014.02.007.

[110] Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drabløs F, et al. Gene regulation. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science. 2015 Feb;347(6225):1010–1014. Available from: http://dx.doi.org/10.1126/science.1259418.

[111] Kowalczyk MS, Hughes JR, Garrick D, Lynch MD, Sharpe JA, Sloane-Stanley JA, et al. Intragenic enhancers act as alternative promoters. Molecular cell. 2012 Feb;45:447–458.

[112] Ørom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, et al. Long noncoding RNAs with enhancer-like function in human cells. Cell. 2010 Oct;143:46–58.

[113] Di Ruscio A, Ebralidze AK, Benoukraf T, Amabile G, Goff LA, Terragni J, et al. DNMT1-interacting RNAs block gene-specific DNA methylation. Nature. 2013 Nov;503:371–376.

[114] Yang YW, Flynn RA, Chen Y, Qu K, Wan B, Wang KC, et al. Essential role of lncRNA binding for WDR5 maintenance of active chromatin and embryonic stem cell pluripotency. eLife. 2014 Feb;3:e02046. Original DateCompleted: 20140213.

[115] Su X, Malouf GG, Chen Y, Zhang J, Yao H, Valero V, et al. Comprehensive analysis of long non-coding RNAs in human breast cancer clinical subtypes. Oncotarget. 2014 Oct;5:9864–9876.

[116] Chen H, Du G, Song X, Li L. Non-coding Transcripts from Enhancers: New Insights into Enhancer Activity and Gene Expression Regulation. Genomics, proteomics & bioinformatics. 2017 Jun;15:201–207.

[117] Outchkourov NS, Muiño JM, Kaufmann K, van Ijcken WFJ, Groot Koerkamp MJ, van Leenen D, et al. Balancing of histone H3K4 methylation states by the Kdm5c/SMCX histone demethylase modulates promoter and enhancer function. Cell Rep. 2013 Apr;3(4):1071–1079. Available from: http://dx.doi.org/10.1016/j.celrep.2013.02.030.

[118] Fong YW, Zhou Q. Stimulatory effect of splicing factors on transcriptional elongation. Nature. 2001;414:929–933.

[119] Austenaa LMI, Barozzi I, Simonatto M, Masella S, Della Chiara G, Ghisletti S, et al. Transcription of Mammalian cis-Regulatory Elements Is Restrained by Actively Enforced Early Termination. Mol Cell. 2015 Nov;60(3):460–474. Available from: http://dx.doi.org/10.1016/j.molcel.2015.09.018.

[120] Meng FL, Du Z, Federation A, Hu J, Wang Q, Kieffer-Kwon KR, et al. Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. Cell. 2014 Dec;159:1538–1548.

[121] Andersson R, Sandelin A, Danko CG. A unified architecture of transcriptional regulatory elements. Trends Genet. 2015 Aug;31(8):426–433. Available from: http://dx.doi.org/10.1016/j.tig.2015.05.007.

[122] Andersson R. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. Bioessays. 2015 Mar;37(3):314–323. Available from: http://dx.doi.org/10.1002/bies.201400162.

[123] Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. Nature reviews Molecular cell biology. 2018 Oct;19:621–637.

[124] Lim LWK, Chung HH, Chong YL, Lee NK. A survey of recently emerged genome-wide computational enhancer predictor tools. Computational biology and chemistry. 2018 Jun;74:132–141.

[125] Ramisch A, Heinrich V, Glaser LV, Fuchs A, Yang X, Benner P, et al. CRUP: a comprehensive framework to predict condition-specific regulatory units. Genome biology. 2019 Nov;20:227.

[126] Benton ML, Talipineni SC, Kostka D, Capra JA. Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function. BMC genomics. 2019 Jun;20:511.

[127] Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, et al. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. Genome research. 2016 Nov;.

[128] Li Y, Okuno Y, Zhang P, Radomska HS, Chen Hm, Iwasaki H, et al. Regulation of the PU. 1 gene by distal elements. Blood. 2001;98(10):2958–2965.

[129] Rosenbauer F, Wagner K, Kutok JL, Iwasaki H, Le Beau MM, Okuno Y, et al. Acute myeloid leukemia induced by graded reduction of a lineage-specific transcription factor, PU. 1. Nature genetics. 2004;36(6):624.

[130] Rosenbauer F, Owens BM, Yu L, Tumang JR, Steidl U, Kutok JL, et al. Lymphoid cell growth and transformation are suppressed by a key regulatory element of the gene encoding PU.1. Nature genetics. 2006 Jan;38:27–37.

[131] Leddin M, Perrod C, Hoogenkamp M, Ghani S, Assi S, Heinz S, et al. Two distinct auto-regulatory loops operate at the PU. 1 locus in B cells and myeloid cells. Blood. 2011;117(10):2827–2838.

[132] Osterwalder M, Barozzi I, Tissières V, Fukuda-Yuzawa Y, Mannion BJ, Afzal SY, et al. Enhancer redundancy provides phenotypic robustness in mammalian development. Nature. 2018 Feb;554:239–243.

[133] Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet. 2014 Apr;15(4):272–286. Available from: http://dx.doi.org/10.1038/nrg3682.

[134] Whitaker JW, Nguyen TT, Zhu Y, Wildberg A, Wang W. Computational schemes for the prediction and annotation of enhancers from epigenomic assays. Methods. 2015 Jan;72:86–94. Available from: http://dx.doi.org/10.1016/j.ymeth.2014.10.008.

[135] Stewart AF, Reik A, Schütz G. A simpler and better method to cleave chromatin with DNase 1 for hypersensitive site analyses. Nucleic acids research. 1991 Jun;19:3157.

[136] Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, et al. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome research. 2006 Jan;16:123–131.

[137] Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012 Sep;489:75–82.

[138] Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome research. 2007 Jun;17:877–885.

[139] Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013 Dec;10(12):1213–1218. Available from: http://dx.doi.org/10.1038/nmeth.2688.

[140] Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015 Jul;523:486–490.

[141] Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science (New York, NY). 2015 May;348:910–914.

[142] Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007 May;129(4):823–837. Available from: http://dx.doi.org/10.1016/j.cell.2007.05.009.

[143] Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science (New York, NY). 2007 Jun;316:1497–1502.

[144] Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. 2007 Aug;448:553–560.

[145] Zentner GE, Tesar PJ, Scacheri PC. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome research. 2011 Aug;21:1273–1283.

[146] Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007 Mar;39(3):311–318. Available from: http://dx.doi.org/10.1038/ng1966.

[147] Pekowska A, Benoukraf T, Zacarias-Cabeza J, Belhocine M, Koch F, Holota H, et al. H3K4 trimethylation provides an epigenetic signature of active enhancers. EMBO J. 2011 Oct;30(20):4198–4210. Available from: http://dx.doi.org/10.1038/emboj.2011.295.

[148] Koch F, Andrau JC. Initiating RNA polymerase II and TIPs as hallmarks of enhancer activity and tissue-specificity. Transcription. 2011;2(6):263–268. Available from: http://dx.doi.org/10.4161/trns.2.6.18747.

[149] Rickels R, Herz HM, Sze CC, Cao K, Morgan MA, Collings CK, et al. Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. Nature genetics. 2017 Nov;49:1647–1653.

[150] Rhee HS, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell. 2011 Dec;147:1408–1419.

[151] Lara-Astiaso D, Weiner A, Lorenzo-Vivas E, Zaretsky I, Jaitin DA, David E, et al. Chromatin state dynamics during blood formation. Science. 2014 Aug;345(6199):943–949. Available from: http://dx.doi.org/10.1126/science.1256271.

[152] Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, Weitz DA, et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nature biotechnology. 2015 Nov;33:1165–1172.

[153] Bornstein C, Winter D, Barnett-Itzhaki Z, David E, Kadri S, Garber M, et al. A negative feedback loop of transcription factors specifies alternative dendritic cell chromatin States. Mol Cell. 2014 Dec;56(6):749–762. Available from: http://dx.doi.org/10.1016/j.molcel.2014.10.014.

[154] Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 2009 Feb;457(7231):854–858. Available from: http://dx.doi.org/10.1038/nature07730.

[155] Aranda-Orgilles B, Saldaña-Meyer R, Wang E, Trompouki E, Fassl A, Lau S, et al. MED12 Regulates HSC-Specific Enhancers Independently of Mediator Kinase Activity to Control Hematopoiesis. Cell stem cell. 2016 Aug;.

[156] Lickwar CR, Mueller F, Hanlon SE, McNally JG, Lieb JD. Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. Nature. 2012 Apr;484:251–255.

[157] Teytelman L, Thurtle DM, Rine J, van Oudenaarden A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. Proceedings of the National Academy of Sciences of the United States of America. 2013 Nov;110:18602–18607.

[158] Gatehouse J, Thompson AJ. Nuclear "run-on" transcription assays. Methods in molecular biology (Clifton, NJ). 1995;49:229–238.

[159] Carninci P, Kvam C, Kitamura A, Ohsumi T, Okazaki Y, Itoh M, et al. High-efficiency full-length cDNA cloning by biotinylated CAP trapper. Genomics. 1996 Nov;37(3):327–336. Available from: http://dx.doi.org/10.1006/geno.1996.0567.

[160] Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A. 2003 Dec;100(26):15776–15781. Available from: http://dx.doi.org/10.1073/pnas.2136655100.

[161] Nagari A, Murakami S, Malladi VS, Kraus WL. In: Ørom UA, editor. Computational Approaches for Mining GRO-Seq Data to Identify and Characterize Active Enhancers. New York, NY: Springer New York; 2017. p. 121–138. Available from: https://doi.org/10.1007/978-1-4939-4035-6_10.

[162] Takahashi H, Lassmann T, Murata M, Carninci P. 5′ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. Nat Protoc. 2012 Mar;7(3):542–561. Available from: http://dx.doi.org/10.1038/nprot.2012.005.

[163] Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science (New York, NY). 2010 May;328:1036–1040.

[164] Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, et al. ChIP-Seq identification of weakly conserved heart enhancers. Nature genetics. 2010 Sep;42:806–810.

[165] Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, et al. Enhancer Evolution across 20 Mammalian Species. Cell. 2015 Jan;160(3):554–566. Available from: http://dx.doi.org/10.1016/j.cell.2015.01.006.

[166] Long HK, Prescott SL, Wysocka J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell. 2016 Nov;167:1170–1187.

[167] Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, et al. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell. 2006 Jan;124:47–59.

[168] Fletez-Brant C, Lee D, McCallion AS, Beer MA. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. Nucleic acids research. 2013 Jul;41:W544–W556.

[169] Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010 May;38(4):576–589. Available from: http://dx.doi.org/10.1016/j.molcel.2010.05.004.

[170] Inoue F, Ahituv N. Decoding enhancers using massively parallel reporter assays. Genomics. 2015 Sep;106:159–164.

[171] Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science (New York, NY). 2013 Mar;339:1074–1077.

[172] Vanhille L, Griffon A, Maqbool MA, Zacarias-Cabeza J, Dao LTM, Fernandez N, et al. High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. Nat Commun. 2015;6:6905. Available from: http://dx.doi.org/10.1038/ncomms7905.

[173] Liu Y, Yu S, Dhiman VK, Brunetti T, Eckart H, White KP. Functional assessment of human enhancer activities using whole-genome STARR-sequencing. Genome biology. 2017 Nov;18:219.

[174] Lemp NA, Hiraoka K, Kasahara N, Logg CR. Cryptic transcripts from a ubiquitous plasmid origin of replication confound tests for cis-regulatory function. Nucleic acids research. 2012 Aug;40:7280–7290.

[175] Muerdter F, Boryń ŁM, Woodfin AR, Neumayr C, Rath M, Zabidi MA, et al. Resolving systematic errors in widely used enhancer activity assays in human cells. Nature methods. 2018 Feb;15:141–149.

[176] Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proceedings of the National Academy of Sciences. 2012;109(47):19498–19503.

[177] Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, et al. Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. Cell. 2016 Jun;165:1519–1529.

[178] White MA. Understanding how cis-regulatory function is encoded in DNA sequence using massively parallel reporter assays and designed sequences. Genomics. 2015 Sep;106:165–170.

[179] Zeigler RD, Cohen BA. Discrimination between thermodynamic models of cis-regulation using transcription factor occupancy data. Nucleic acids research. 2014 Feb;42:2224–2234.

[180] Sherman MS, Cohen BA. Thermodynamic state ensemble models of cis-regulation. PLoS computational biology. 2012;8:e1002407.

[181] de Souza FS, Franchini LF, Rubinstein M. Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? Molecular biology and evolution. 2013;30(6):1239–1251.

[182] Planello AC, Singhania R, Kron KJ, Bailey SD, Roulois D, Lupien M, et al. Pre-neoplastic epigenetic disruption of transcriptional enhancers in chronic inflammation. Oncotarget. 2016 Feb;Available from: http://dx.doi.org/10.18632/oncotarget.7513.

[183] Kioussis D, Vanin E, deLange T, Flavell RA, Grosveld FG. Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. Nature. 1983;306:662–666.

[184] Taramelli R, Kioussis D, Vanin E, Bartram K, Groffen J, Hurst J, et al. Gamma delta beta-thalassaemias 1 and 2 are the result of a 100 kbp deletion in the human beta-globin cluster. Nucleic acids research. 1986 Sep;14:7017–7029.

[185] Harteveld CL, Voskamp A, Phylipsen M, Akkermans N, den Dunnen JT, White SJ, et al. Nine unknown rearrangements in 16p13.3 and 11p15.4 causing alpha- and beta-thalassaemia characterised by high resolution multiplex ligation-dependent probe amplification. Journal of medical genetics. 2005 Dec;42:922–931.

[186] Smemo S, Campos LC, Moskowitz IP, Krieger JE, Pereira AC, Nobrega MA. Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. Human molecular genetics. 2012 Jul;21:3255–3263.

[187] Emison ES, McCallion AS, Kashuk CS, Bush RT, Grice E, Lin S, et al. A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. Nature. 2005 Apr;434:857–863.

[188] Kikuchi M, Hara N, Hasegawa M, Miyashita A, Kuwano R, Ikeuchi T, et al. Enhancer variants associated with Alzheimer's disease affect gene expression via chromatin looping. BMC medical genomics. 2019 Sep;12:128.

[189] Heshka JT, Palleschi C, Howley H, Wilson B, Wells PS. A systematic review of perceived risks, psychological and behavioral impacts of genetic testing. Genetics in medicine : official journal of the American College of Medical Genetics. 2008 Jan;10:19–32.

[190] Lin CY, Erkek S, Tong Y, Yin L, Federation AJ, Zapatka M, et al. Active medulloblastoma enhancers reveal subgroup-specific cellular origins. Nature. 2016 Feb;530:57–62.

[191] Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nature genetics. 2009 Aug;41:882–884.

[192] Wasserman NF, Aneas I, Nobrega MA. An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. Genome research. 2010 Sep;20:1191–1197.

[193] Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nature genetics. 2009 Aug;41:885–890.

[194] Herz HM, Hu D, Shilatifard A. Enhancer malfunction in cancer. Molecular cell. 2014;53(6):859–866.

[195] Smith E, Shilatifard A. Enhancer biology and enhanceropathies. Nat Struct Mol Biol. 2014 Mar;21(3):210–219. Available from: http://dx.doi.org/10.1038/nsmb.2784.

[196] Sur I, Taipale J. The role of enhancers in cancer. Nature reviews Cancer. 2016 Aug;16:483–493.

[197] Ulirsch JC, Lareau CA, Bao EL, Ludwig LS, Guo MH, Benner C, et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. Nature genetics. 2019 Mar;.

[198] Bresnick EH, Johnson KD. Blood disease-causing and -suppressing transcriptional enhancers: general principles and and GATA2 mechanisms. Blood advances. 2019 Jul;3:2045–2056.

[199] Zuber J, Shi J, Wang E, Rappaport AR, Herrmann H, Sison Ea, et al. RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia. Nature. 2011 Oct;478(7370):524–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21814200.

[200] Shi J, Whyte WA, Zepeda-Mendoza CJ, Milazzo JP, Shen C, Roe JS, et al. Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation. Genes & development. 2013 Dec;27:2648–2662.

[201] Metcalf D, Dakic A, Mifsud S, Di Rago L, Wu L, Nutt S. Inactivation of PU.1 in adult mice leads to the development of myeloid leukemia. Proceedings of the National Academy of Sciences of the United States of America. 2006 Jan;103:1486–1491.

[202] Will B, Vogler TO, Narayanagari S, Bartholdy B, Todorova TI, da Silva Ferreira M, et al. Minimal PU.1 reduction induces a preleukemic state and promotes development of acute myeloid leukemia. Nat Med. 2015 Oct;21(10):1172–1181. Available from: http://dx.doi.org/10.1038/nm.3936.

[203] Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nat Genet. 2016 Aug;Available from: http://dx.doi.org/10.1038/ng.3646.

[204] Mansour MR, Abraham BJ, Anders L, Berezovskaya A, Gutierrez A, Durbin AD, et al. Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Science. 2014 Dec;346(6215):1373–1377. Available from: http://dx.doi.org/10.1126/science.1259037.

[205] Vahedi G, Kanno Y, Furumoto Y, Jiang K, Parker SCJ, Erdos MR, et al. Super-enhancers delineate disease-associated regulatory nodes in T cells. Nature. 2015 Feb;Available from: http://dx.doi.org/10.1038/nature14154.

[206] Huang Y, Mouttet B, Warnatz HJ, Risch T, Rietmann F, Frommelt F, et al. The Leukemogenic TCF3-HLF Complex Rewires Enhancers Driving Cellular Identity and Self-Renewal Conferring EP300 Vulnerability. Cancer cell. 2019 Nov;.

[207] Bahr C, von Paleske L, Uslu VV, Remeseiro S, Takayama N, Ng SW, et al. A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. Nature. 2018 Jan;553:515–520.

[208] Arber DA, Orazi A, Hasserjian R, Thiele J, Borowitz MJ, Le Beau MM, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood. 2016 May;127:2391–2405.

[209] Yamazaki H, Suzuki M, Otsuki A, Shimizu R, Bresnick EH, Engel JD, et al. A remote GATA2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating EVI1 expression. Cancer Cell. 2014 Apr;25(4):415–427. Available from: http://dx.doi.org/10.1016/j.ccr.2014.02.008.

[210] Gröschel S, Sanders MA, Hoogenboezem R, de Wit E, Bouwman BAM, Erpelinck C, et al. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. Cell. 2014 Apr;157(2):369–381. Available from: http://dx.doi.org/10.1016/j.cell.2014.02.019.

[211] Bröske AM, Vockentanz L, Kharazi S, Huska MR, Mancini E, Scheller M, et al. DNA methylation protects hematopoietic stem cell multipotency from myeloerythroid restriction. Nat Genet. 2009 Nov;41(11):1207–1215. Available from: http://dx.doi.org/10.1038/ng.463.

[212] Trowbridge JJ, Sinha AU, Zhu N, Li M, Armstrong Sa, Orkin SH. Haploinsufficiency of Dnmt1 impairs leukemia stem cell function through derepression of bivalent chromatin domains. Genes & development. 2012 Feb;26(4):344–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22345515.

[213] Krivtsov AV, Twomey D, Feng Z, Stubbs MC, Wang Y, Faber J, et al. Transformation from committed progenitor to leukaemia stem cell initiated by MLL-AF9. Nature. 2006 Aug;442(7104):818–822. Available from: http://dx.doi.org/10.1038/nature04980.

[214] Somervaille TCP, Cleary ML. Identification and characterization of leukemia stem cells in murine MLL-AF9 acute myeloid leukemia. Cancer Cell. 2006 Oct;10(4):257–68. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17045204.

[215] Jeong M, Sun D, Luo M, Huang Y, Challen GA, Rodriguez B, et al. Large conserved domains of low DNA methylation maintained by Dnmt3a. Nat Genet. 2014 Jan;46(1):17–23. Available from: http://dx.doi.org/10.1038/ng.2836.

[216] Challen GA, Sun D, Mayle A, Jeong M, Luo M, Rodriguez B, et al. Dnmt3a and Dnmt3b have overlapping and distinct functions in hematopoietic stem cells. Cell stem cell. 2014 Sep;15:350–364.

[217] Wiehle L, Raddatz G, Musch T, Dawlaty MM, Jaenisch R, Lyko F, et al. Tet1 and Tet2 Protect DNA Methylation Canyons against Hypermethylation. Molecular and cellular biology. 2016 Feb;36:452–461.

[218] Schulze I, Rohde C, Scheller-Wendorff M, Bäumer N, Krause A, Herbst F, et al. Increased DNA methylation of Dnmt3b targets impairs leukemogenesis. Blood. 2016 Mar;127:1575–1586.

[219] Rulands S, Lee HJ, Clark SJ, Angermueller C, Smallwood SA, Krueger F, et al. Genome-Scale Oscillations in DNA Methylation during Exit from Pluripotency. Cell systems. 2018 Jul;7:63–76.e12.

[220] Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009 Nov;462(7271):315–322. Available from: http://dx.doi.org/10.1038/nature08514.

[221] Schroeder DI, LaSalle JM. How has the study of the human placenta aided our understanding of partially methylated genes? Epigenomics. 2013 Dec;5(6):645–654. Available from: http://dx.doi.org/10.2217/epi.13.62.

[222] Gaidatzis D, Burger L, Murr R, Lerch A, Dessus-Babus S, Schübeler D, et al. DNA sequence explains seemingly disordered methylation levels in partially methylated domains of Mammalian genomes. PLoS Genet. 2014 Feb;10(2):e1004143. Available from: http://dx.doi.org/10.1371/journal.pgen.1004143.

[223] Burger L, Gaidatzis D, Schübeler D, Stadler MB. Identification of active regulatory regions from DNA methylation data. Nucleic Acids Res. 2013 Sep;41(16):e155. Available from: http://dx.doi.org/10.1093/nar/gkt599.

[224] Kuan PF, Wang S, Zhou X, Chu H. A statistical framework for Illumina DNA methylation arrays. Bioinformatics. 2010 Nov;26(22):2849–2855. Available from: http://dx.doi.org/10.1093/bioinformatics/btq553.

[225] Hebestreit K, Dugas M, Klein HU. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. Bioinformatics. 2013 Jul;29(13):1647–1653. Available from: http://dx.doi.org/10.1093/bioinformatics/btt263.

[226] Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, et al. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. PLoS One. 2013;8(12):e81148. Available from: http://dx.doi.org/10.1371/journal.pone.0081148.

[227] Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature. 2008 Jun;453(7197):948–951. Available from: http://dx.doi.org/10.1038/nature06947.

[228] Smithson M, Verkuilen J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychological methods. 2006;11(1):54.

[229] Kumaraswamy P. A generalized probability density function for double-bounded random processes. Journal of Hydrology. 1980;46(1-2):79–88. Available from: http://www.sciencedirect.com/science/article/pii/0022169480900360.

[230] Jones M. Kumaraswamys distribution: A beta-type distribution with some tractability advantages. Statistical Methodology. 2009;6(1):70–81.

[231] Meuleman W, Peric-Hupkes D, Kind J, Beaudry JB, Pagie L, Kellis M, et al. Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. Genome Res. 2013 Feb;23(2):270–280. Available from: http://dx.doi.org/10.1101/gr.141028.112.

[232] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar;26(6):841–842. Available from: http://dx.doi.org/10.1093/bioinformatics/btq033.

[233] Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011 Dec;480(7378):490–495. Available from: http://dx.doi.org/10.1038/nature10716.

[234] Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LTY, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013 Aug;500(7463):477–481. Available from: http://dx.doi.org/10.1038/nature12433.

[235] Thiery JP, Macaya G, Bernardi G. An analysis of eukaryotic genomes by density gradient centrifugation. Journal of molecular biology. 1976 Nov;108:219–235.

[236] Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, et al. The mosaic genome of warm-blooded vertebrates. Science (New York, NY). 1985 May;228:953–958.

[237] Hurst LD, Merchant AR. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proceedings Biological sciences. 2001 Mar;268:493–497.

[238] Eyre-Walker A, Hurst LD. The evolution of isochores. Nature reviews Genetics. 2001 Jul;2:549–555.

[239] Ream RA, Johns GC, Somero GN. Base compositions of genes encoding alpha-actin and lactate dehydrogenase-A from differently adapted vertebrates show no temperature-adaptive variation in G + C content. Molecular biology and evolution. 2003 Jan;20:105–110.

[240] Cohen N, Dagan T, Stone L, Graur D. GC composition of the human genome: in search of isochores. Molecular biology and evolution. 2005 May;22:1260–1272.

[241] Romiguier J, Ranwez V, Douzery EJP, Galtier N. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. Genome research. 2010 Aug;20:1001–1009.

[242] Jabbari K, Bernardi G. An Isochore Framework Underlies Chromatin Architecture. PloS one. 2017;12:e0168023.

[243] Oliver JL, Carpena P, Hackenberg M, Bernaola-Galván P. IsoFinder: computational prediction of isochores in genome sequences. Nucleic acids research. 2004 Jul;32:W287–W292.

[244] Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nat Genet. 2012 Jan;44(1):40–46. Available from: http://dx.doi.org/10.1038/ng.969.

[245] Timp W, Bravo HC, McDonald OG, Goggins M, Umbricht C, Zeiger M, et al. Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. Genome Med. 2014;6(8):61. Available from: http://dx.doi.org/10.1186/s13073-014-0061-y.

[246] Wilson NK, Schoenfelder S, Hannah R, Sánchez Castillo M, Schütte J, Ladopoulos V, et al. Integrated genome-scale analysis of the transcriptional regulatory landscape in a blood stem/progenitor cell model. Blood. 2016 Mar;127(13):e12–e23. Available from: http://dx.doi.org/10.1182/blood-2015-10-677393.

[247] Wilson NK, Foster SD, Wang X, Knezevic K, Schütte J, Kaimakis P, et al. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. Cell stem cell. 2010 Oct;7:532–544.

[248] Calero-Nieto FJ, Ng FS, Wilson NK, Hannah R, Moignard V, Leal-Cervantes AI, et al. Key regulators control distinct transcriptional programmes in blood progenitor and mast cells. The EMBO journal. 2014 Jun;33:1212–1226.

[249] Schütte J, Wang H, Antoniou S, Jarratt A, Wilson NK, Riepsaame J, et al. An experimentally validated network of nine haematopoietic transcription factors reveals mechanisms of cell state stability. eLife. 2016 Feb;5:e11469.

[250] Kruse K, Hug CB, Hernández-Rodríguez B, Vaquerizas JM. TADtool: visual parameter identification for TAD-calling algorithms. Bioinformatics (Oxford, England). 2016 Oct;32:3190–3192.

[251] Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature. 2015 Jul;523:240–244.

[252] Vockentanz L, Bröske AM, Rosenbauer F. Uncovering a unique role for DNA methylation in hematopoietic and leukemic stem cells. Cell cycle (Georgetown, Tex). 2010 Feb;9(4):640–1. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20107327.

[253] Kahles A, Lehmann KV, Toussaint NC, Hüser M, Stark SG, Sachsenberg T, et al. Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer cell. 2018 Aug;34:211–224.e6.

[254] Mendizabal I, Zeng J, Keller TE, Yi SV. Body-hypomethylated human genes harbor extensive intragenic transcriptional activity and are prone to cancer-associated dysregulation. Nucleic acids research. 2017 May;45:4390–4400.

[255] Nanan KK, Ocheltree C, Sturgill D, Mandler MD, Prigge M, Varma G, et al. Independence between pre-mRNA splicing and DNA methylation in an isogenic minigene resource. Nucleic acids research. 2017 Dec;45:12780–12797.

[256] Brocks D, Schmidt CR, Daskalakis M, Jang HS, Shah NM, Li D, et al. DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. Nature genetics. 2017 Jun;.

[257] Cai Y, Tsai HC, Yen RWC, Zhang YW, Kong X, Wang W, et al. Critical threshold levels of DNA methyltransferase 1 are required to maintain DNA methylation across the genome in human cancer cells. Genome research. 2017 Apr;27:533–544.

[258] Faulkner GJ, Forrest ARR, Chalk AM, Schroder K, Hayashizaki Y, Carninci P, et al. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. Genomics. 2008 Mar;91(3):281–288. Available from: http://dx.doi.org/10.1016/j.ygeno.2007.11.003.

[259] Slany RK. The molecular mechanics of mixed lineage leukemia. Oncogene. 2016;35(40):5215–5223.

[260] Adra CN, Boer PH, McBurney MW. Cloning and expression of the mouse pgk-1 gene and the nucleotide sequence of its promoter. Gene. 1987;60(1):65–74.

[261] Tybulewicz VL, Crawford CE, Jackson PK, Bronson RT, Mulligan RC. Neonatal lethality and lymphopenia in mice with a homozygous disruption of the c-abl proto-oncogene. Cell. 1991 Jun;65(7):1153–1163.

[262] Lei H, Oh SP, Okano M, Jüttermann R, Goss KA, Jaenisch R, et al. De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. Development. 1996 Oct;122(10):3195–3205.

[263] Yenofsky RL, Fine M, Pellow JW. A mutant neomycin phosphotransferase II gene reduces the resistance of transformants to antibiotic selection pressure. Proceedings of the National Academy of Sciences of the United States of America. 1990 May;87:3435–3439.

[264] Tucker KL, Talbot D, Lee MA, Leonhardt H, Jaenisch R. Complementation of methylation deficiency in embryonic stem cells by a DNA methyltransferase minigene. Proc Natl Acad Sci U S A. 1996 Nov;93(23):12920–12925.

[265] Gaudet F, Hodgson JG, Eden A, Jackson-Grusby L, Dausman J, Gray JW, et al. Induction of tumors in mice by genomic hypomethylation. Science. 2003 Apr;300(5618):489–492. Available from: http://dx.doi.org/10.1126/science.1083558.

[266] Kind J, Pagie L, de Vries SS, Nahidiazar L, Dey SS, Bienko M, et al. Genome-wide maps of nuclear lamina interactions in single human cells. Cell. 2015 Sep;163(1):134–147. Available from: http://dx.doi.org/10.1016/j.cell.2015.08.040.

[267] Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome biology. 2013;14:R95.

[268] Zhang ZH, Jhaveri DJ, Marshall VM, Bauer DC, Edson J, Narayanan RK, et al. A comparative study of techniques for differential expression analysis on RNA-Seq data. PloS one. 2014;9:e103207.

[269] Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Briefings in bioinformatics. 2015 Jan;16:59–70.

[270] Tang M, Sun J, Shimizu K, Kadota K. Evaluation of methods for differential expression analysis on multi-group RNA-seq count data. BMC bioinformatics. 2015 Nov;16:361.

[271] Jaakkola MK, Seyednasrollah F, Mehmood A, Elo LL. Comparison of methods to detect differentially expressed genes between single-cell populations. Briefings in bioinformatics. 2017 Sep;18:735–743.

[272] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological). 1995;p. 289–300.

[273] Anders S, Huber W. Differential expression analysis for sequence count data. Genome biology. 2010;11(10):R106.

[274] McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic acids research. 2012 May;40:4288–4297.

[275] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England). 2010 Jan;26:139–140.

[276] Chen Y, Lun ATL, Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000Research. 2016;5:1438.

[277] Chen W, Kumar AR, Hudson Wa, Li Q, Wu B, Staggs Ra, et al. Malignant transformation initiated by Mll-AF9: gene dosage and critical target cells. Cancer cell. 2008 May;13(5):432–40. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2430522&tool=pmcentrez&rendertype=abstract.

[278] Somervaille TCP, Matheny CJ, Spencer GJ, Iwasaki M, Rinn JL, Witten DM, et al. Hierarchical maintenance of MLL myeloid leukemia stem cells employs a transcriptional program shared with embryonic rather than adult stem cells. Cell Stem Cell. 2009 Feb;4(2):129–140. Available from: http://dx.doi.org/10.1016/j.stem.2008.11.015.

[279] Gentles AJ, Plevritis SK, Majeti R, Alizadeh AA. Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloid leukemia. JAMA. 2010 Dec;304(24):2706–2715. Available from: http://dx.doi.org/10.1001/jama.2010.1862.

[280] Eppert K, Takenaka K, Lechman ER, Waldron L, Nilsson B, van Galen P, et al. Stem cell gene expression programs influence clinical outcome in human leukemia. Nat Med. 2011 Sep;17(9):1086–1093. Available from: http://dx.doi.org/10.1038/nm.2415.

[281] Cabezas-Wallscheid N, Eichwald V, de Graaf J, Löwer M, Lehr HA, Kreft A, et al. Instruction of haematopoietic lineage choices, evolution of transcriptional landscapes and cancer stem cell hierarchies derived from an AML1-ETO mouse model. EMBO Mol Med. 2013 Dec;5(12):1804–1820. Available from: http://dx.doi.org/10.1002/emmm.201302661.

[282] Stavropoulou V, Kaspar S, Brault L, Sanders MA, Juge S, Morettini S, et al. MLL-AF9 Expression in Hematopoietic Stem Cells Drives a Highly Invasive AML Expressing EMT-Related Genes Linked to Poor Outcome. Cancer Cell. 2016 Jun;Available from: http://dx.doi.org/10.1016/j.ccell.2016.05.011.

[283] Gupta R, Hong D, Iborra F, Sarno S, Enver T. NOV (CCN3) functions as a regulator of human hematopoietic stem or progenitor cells. Science (New York, NY). 2007 Apr;316:590–593.

[284] Yin XY, Gupta K, Han WP, Levitan ES, Prochownik EV. Mmip-2, a novel RING finger protein that interacts with mad members of the Myc oncoprotein network. Oncogene. 1999 Nov;18(48):6621–6634. Available from: http://dx.doi.org/10.1038/sj.onc.1203097.

[285] Wenda JM, Homolka D, Yang Z, Spinelli P, Sachidanandam R, Pandey RR, et al. Distinct Roles of RNA Helicases MVH and TDRD9 in PIWI Slicing-Triggered Mammalian piRNA Biogenesis and Function. Developmental cell. 2017 Jun;41:623–637.e9.

[286] Benayoun BA, Pollina EA, Ucar D, Mahmoudi S, Karra K, Wong ED, et al. H3K4me3 breadth is linked to cell identity and transcriptional consistency. Cell. 2014 Jul;158(3):673–688. Available from: http://dx.doi.org/10.1016/j.cell.2014.06.027.

[287] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome biology. 2008;9:R137.

[288] Thomas R, Thomas S, Holloway AK, Pollard KS. Features that define the best ChIP-seq peak calling algorithms. Briefings in bioinformatics. 2017 May;18:441–450.

[289] Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics. 2009 Aug;25(15):1952–1958. Available from: http://dx.doi.org/10.1093/bioinformatics/btp340.

[290] Xu S, Grullon S, Ge K, Peng W. Spatial clustering for identification of ChIP-enriched regions (SICER) to map regions of histone methylation patterns in embryonic stem cells. Methods in molecular biology (Clifton, NJ). 2014;1150:97–111.

[291] Steinhauser S, Kurzawa N, Eils R, Herrmann C. A comprehensive comparison of tools for differential ChIP-seq analysis. Briefings in bioinformatics. 2016 Nov;17:953–966.

[292] Cao F, Fang Y, Tan HK, Goh Y, Choy JYH, Koh BTH, et al. Super-Enhancers and Broad H3K4me3 Domains Form Complex Gene Regulatory Circuits Involving Chromatin Interactions. Scientific reports. 2017 May;7:2186.

[293] Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013 Aug;8(8):1494–1512. Available from: http://dx.doi.org/10.1038/nprot.2013.084.

[294] Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic acids research. 2013 May;41:e108.

[295] Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics (Oxford, England). 2014 Apr;30:923–930.

[296] Maretty L, Sibbesen JA, Krogh A. Bayesian transcriptome assembly. Genome biology. 2014;15:501.

[297] Bushnell B. BBMap short read aligner and other bioinformatic tools. Lawrence Berkeley National Laboratory; 2014. Available from: https://sourceforge.net/projects/bbmap/.

[298] Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology. 2015 Mar;33:290–295.

[299] Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature protocols. 2016 Sep;11:1650–1667.

[300] Delas MJ, Sabin LR, Dolzhenko E, Knott SR, Maravilla EM, Jackson BT, et al. lncRNA requirements for mouse acute myeloid leukemia and normal differentiation. Elife. 2017;6.

[301] Mohan M, Lin C, Guest E, Shilatifard A. Licensed to elongate: a molecular mechanism for MLL-based leukaemogenesis. Nature reviews Cancer. 2010 Oct;10:721–728.

[302] Frith MC, Valen E, Krogh A, Hayashizaki Y, Carninci P, Sandelin A. A code for transcription initiation in mammalian genomes. Genome Research. 2008 Jan;18:1–12.

[303] Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome biology. 2015 Jan;16:22.

[304] Ye M, Zhang H, Yang H, Koche R, Staber PB, Cusan M, et al. Hematopoietic Differentiation Is Required for Initiation of Acute Myeloid Leukemia. Cell Stem Cell. 2015 Nov;17(5):611–623. Available from: http://dx.doi.org/10.1016/j.stem.2015.08.011.

[305] George J, Uyar A, Young K, Kuffler L, Waldron-Francis K, Marquez E, et al. Leukaemia cell of origin identified by chromatin landscape of bulk tumour cells. Nature communications. 2016 Jul;7:12166.

[306] Lipka DB, Wang Q, Cabezas-Wallscheid N, Klimmeck D, Weichenhan D, Herrmann C, et al. Identification of DNA methylation changes at cis-regulatory elements during early steps of HSC differentiation using tagmentation-based whole genome bisulfite sequencing. Cell Cycle. 2014;13(22):3476–3487. Available from: http://dx.doi.org/10.4161/15384101.2014.973334.

[307] Schübeler D. Function and information content of DNA methylation. Nature. 2015 Jan;517:321–326.

[308] Hon GC, Rajagopal N, Shen Y, McCleary DF, Yue F, Dang MD, et al. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. Nat Genet. 2013 Oct;45(10):1198–1206. Available from: http://dx.doi.org/10.1038/ng.2746.

[309] Schlesinger F, Smith AD, Gingeras TR, Hannon GJ, Hodges E. De novo DNA demethylation and noncoding transcription define active intergenic regulatory elements. Genome Res. 2013 Oct;23(10):1601–1614. Available from: http://dx.doi.org/10.1101/gr.157271.113.

[310] Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, et al. Dynamic DNA methylation across diverse human cell lines and tissues. Genome Res. 2013 Mar;23(3):555–567. Available from: http://dx.doi.org/10.1101/gr.147942.112.

[311] Sheaffer KL, Kim R, Aoki R, Elliott EN, Schug J, Burger L, et al. DNA methylation is required for the control of stem cell differentiation in the small intestine. Genes Dev. 2014 Mar;28(6):652–664. Available from: http://dx.doi.org/10.1101/gad.230318.113.

[312] Allen MD, Grummitt CG, Hilcenko C, Min SY, Tonkin LM, Johnson CM, et al. Solution structure of the nonmethyl-CpG-binding CXXC domain of the leukaemia-associated MLL histone methyltransferase. The EMBO journal. 2006 Oct;25:4503–4512.

[313] Hu S, Wan J, Su Y, Song Q, Zeng Y, Nguyen HN, et al. DNA methylation presents distinct binding sites for human transcription factors. Elife. 2013;2:e00726. Available from: http://dx.doi.org/10.7554/eLife.00726.

[314] Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science (New York, NY). 2017 May;356.

[315] Kemme CA, Marquez R, Luu RH, Iwahara J. Potential role of DNA methylation as a facilitator of target search processes for transcription factors through interplay with methyl-CpG-binding proteins. Nucleic acids research. 2017 Jul;45:7751–7759.

[316] Vockentanz L. Leukemia Stem Cell Fates are Determined by DNA Methylation Levels [phdthesis]. Mathematisch-Naturwissenschaftlichen Fakultät I der Humboldt-Universität zu Berlin; 2011. Available from: https://edoc.hu-berlin.de/bitstream/handle/18452/16973/vockentanz.pdf.

[317] Hahsler M, Chelluboina S, Hornik K, Buchta C. The arules R-package ecosystem: analyzing interesting patterns from large transaction data sets. Journal of Machine Learning Research. 2011;12(Jun):2021–2025.

[318] Ha N, Polychronidou M, Lohmann I. COPS: detecting co-occurrence and spatial arrangement of transcription factor binding motifs in genome-wide datasets. PloS one. 2012;7:e52055.

[319] Navarro C, Lopez FJ, Cano C, Garcia-Alcalde F, Blanco A. CisMiner: genome-wide in-silico cis-regulatory module prediction by fuzzy itemset mining. PloS one. 2014;9:e108065.

[320] Agrawal R, Srikant R, et al. Fast algorithms for mining association rules. In: Proc. 20th int. conf. very large data bases, VLDB. vol. 1215; 1994. p. 487–499.

[321] Fujino T, Yamazaki Y, Largaespada DA, Jenkins NA, Copeland NG, Hirokawa K, et al. Inhibition of myeloid differentiation by Hoxa9, Hoxb8, and Meis homeobox genes. Experimental hematology. 2001;29(7):856–863.

[322] Adamaki M, Lambrou GI, Athanasiadou A, Vlahopoulos S, Papavassiliou AG, Moschovi M. HOXA9 and MEIS1 gene overexpression in the diagnosis of childhood acute leukemias: Significant correlation with relapse and overall survival. Leukemia research. 2015;39(8):874–882.

[323] Carter B, Milella M, Tsao T, McQueen T, Schober W, Hu W, et al. Regulation and targeting of antiapoptotic XIAP in acute myeloid leukemia. Leukemia. 2003;17(11):2081–2089.

[324] Carter BZ, Gronda M, Wang Z, Welsh K, Pinilla C, Andreeff M, et al. Small-molecule XIAP inhibitors derepress downstream effector caspases and induce apoptosis of acute myeloid leukemia cells. Blood. 2005;105(10):4043–4050.

[325] Fullard JF, Rahman S, Roussos P. Genetic Variation in Long-Range Enhancers. Current topics in behavioral neurosciences. 2019 Aug;.

[326] Hait TA, Amar D, Shamir R, Elkon R. FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map. Genome biology. 2018 May;19:56.

[327] Walter C, Schuetzmann D, Rosenbauer F, Dugas M. Benchmarking of 4C-seq pipelines based on real and simulated data. Bioinformatics (Oxford, England). 2019 May;.

[328] Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome research. 2012 Sep;22:1680–1688.

[329] Ito Y, Nativio R, Murrell A. Induced DNA demethylation can reshape chromatin topology at the IGF2-H19 locus. Nucleic acids research. 2013 May;41:5290–5302.

[330] Kang JY, Song SH, Yun J, Jeon MS, Kim HP, Han SW, et al. Disruption of CTCF/cohesin-mediated high-order chromatin structures by DNA methylation downregulates PTGS2 expression. Oncogene. 2015 Nov;34:5677–5684.

[331] Whalen S, Truty RM, Pollard KS. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat Genet. 2016 Apr;Available from: http://dx.doi.org/10.1038/ng.3539.

[332] Koeppel M, van Heeringen SJ, Smeenk L, Navis AC, Janssen-Megens EM, Lohrum M. The novel p53 target gene IRF2BP2 participates in cell survival during the p53 stress response. Nucleic acids research. 2009 Feb;37:322–335.

[333] Stadhouders R, Cico A, Stephen T, Thongjuea S, Kolovos P, Baymaz HI, et al. Control of developmentally primed erythroid genes by combinatorial co-repressor actions. Nature communications. 2015 Nov;6:8893.

[334] Wang L, Gao S, Wang H, Xue C, Liu X, Yuan H, et al. Interferon regulatory factor 2 binding protein 2b regulates neutrophil versus macrophage fate during zebrafish definitive myelopoiesis. Haematologica. 2019 May;.

[335] Jovanovic JV, Chillón MC, Vincent-Fabert C, Dillon R, Voisset E, Gutiérrez NC, et al. The cryptic IRF2BP2-RARA fusion transforms hematopoietic stem/progenitor cells and induces retinoid-sensitive acute promyelocytic leukemia. Leukemia. 2017 Mar;31:747–751.

[336] Sun H, Lesche R, Li DM, Liliental J, Zhang H, Gao J, et al. PTEN modulates cell cycle progression and cell survival by regulating phosphatidylinositol 3,4,5,-trisphosphate and Akt/protein kinase B signaling pathway. Proceedings of the National Academy of Sciences of the United States of America. 1999 May;96:6199–6204.

[337] Hu T, Li C, Zhang Y, Wang L, Peng L, Cheng H, et al. Phosphoinositide-dependent kinase 1 regulates leukemia stem cell maintenance in MLL-AF9-induced murine acute myeloid leukemia. Biochemical and biophysical research communications. 2015 Apr;459:692–698.

[338] Sandhöfer N, Metzeler KH, Rothenberg M, Herold T, Tiedt S, GroiSS V, et al. Dual PI3K/mTOR inhibition shows antileukemic activity in MLL-rearranged acute myeloid leukemia. Leukemia. 2015 Apr;29:828–838.

[339] Alimonti A, Carracedo A, Clohessy JG, Trotman LC, Nardella C, Egia A, et al. Subtle variations in Pten dose determine cancer susceptibility. Nature genetics. 2010 May;42:454–458.

[340] Kwabi-Addo B, Giri D, Schmidt K, Podsypanina K, Parsons R, Greenberg N, et al. Haploinsufficiency of the Pten tumor suppressor gene promotes prostate cancer progression. Proceedings of the National Academy of Sciences of the United States of America. 2001 Sep;98:11563–11568.

[341] Trotman LC, Niki M, Dotan ZA, Koutcher JA, Di Cristofano A, Xiao A, et al. Pten dose dictates cancer progression in the prostate. PLoS biology. 2003 Dec;1:E59.

[342] Parisotto M, Grelet E, El Bizri R, Dai Y, Terzic J, Eckert D, et al. PTEN deletion in luminal cells of mature prostate induces replication stress and senescence in vivo. The Journal of experimental medicine. 2018 Jun;215:1749–1763.

[343] Thivierge C, Tseng HW, Mayya VK, Lussier C, Gravel SP, Duchaine TF. Alternative polyadenylation confers Pten mRNAs stability and resistance to microRNAs. Nucleic acids research. 2018 Nov;46:10340–10352.

[344] Carracedo A, Alimonti A, Pandolfi PP. PTEN level in tumor suppression: how much is too little? Cancer research. 2011 Feb;71:629–633.

[345] Bermúdez Brito M, Goulielmaki E, Papakonstanti EA. Focus on PTEN Regulation. Frontiers in oncology. 2015;5:166.

[346] Stam RW, Den Boer ML, Schneider P, de Boer J, Hagelstein J, Valsecchi MG, et al. Association of high-level MCL-1 expression with in vitro and in vivo prednisone resistance in MLL-rearranged infant acute lymphoblastic leukemia. Blood. 2010 Feb;115:1018–1025.

[347] Nonami A, Kato R, Taniguchi K, Yoshiga D, Taketomi T, Fukuyama S, et al. Spred-1 negatively regulates interleukin-3-mediated ERK/mitogen-activated protein (MAP) kinase activation in hematopoietic cells. The Journal of biological chemistry. 2004 Dec;279:52543–52551.

[348] Marschalek R. Mechanisms of leukemogenesis by MLL fusion proteins. British journal of haematology. 2011 Jan;152:141–154.

[349] Pasmant E, Ballerini P, Lapillonne H, Perot C, Vidaud D, Leverger G, et al. SPRED1 disorder and predisposition to leukemia in children. Blood. 2009 Jul;114:1131.

[350] Batz C, Hasle H, Bergsträsser E, van den Heuvel-Eibrink MM, Zecca M, Niemeyer CM, et al. Does SPRED1 contribute to leukemogenesis in juvenile myelomonocytic leukemia (JMML)? Blood. 2010 Mar;115:2557–2558.

[351] Treiber N, Treiber T, Zocher G, Grosschedl R. Structure of an Ebf1:DNA complex reveals unusual DNA recognition and structural homology with Rel proteins. Genes & development. 2010 Oct;24:2270–2275.

[352] Teye EK, Sido A, Xin P, Finnberg NK, Gokare P, Kawasawa YI, et al. PIGN gene expression aberration is associated with genomic instability and leukemic progression in acute myeloid leukemia with myelodysplastic features. Oncotarget. 2017 May;8:29887–29905.

[353] Metzakopian E, Strong A, Iyer V, Hodgkins A, Tzelepis K, Antunes L, et al. Enhancing the genome editing toolbox: genome wide CRISPR arrayed libraries. Scientific reports. 2017 May;7:2244.

[354] Matsuo Y, MacLeod RA, Uphoff CC, Drexler HG, Nishizaki C, Katayama Y, et al. Two acute monocytic leukemia (AML-M5a) cell lines (MOLM-13 and MOLM-14) with interclonal phenotypic heterogeneity showing MLL-AF9 fusion resulting from an occult chromosome insertion, ins(11;9)(q23;p22p23). Leukemia. 1997 Sep;11:1469–1477.

[355] Köhler A, Hurt E. Gene regulation by nucleoporins and links to cancer. Molecular cell. 2010 Apr;38:6–15.

[356] Griffis ER, Craige B, Dimaano C, Ullman KS, Powers MA. Distinct functional domains within nucleoporins Nup153 and Nup98 mediate transcription-dependent mobility. Molecular biology of the cell. 2004 Apr;15:1991–2002.

[357] Hou C, Corces VG. Nups take leave of the nuclear envelope to regulate transcription. Cell. 2010 Feb;140:306–308.

[358] Kalverda B, Pickersgill H, Shloma VV, Fornerod M. Nucleoporins directly stimulate expression of developmental and cell-cycle genes inside the nucleoplasm. Cell. 2010 Feb;140:360–371.

[359] Nakamura T, Largaespada DA, Lee MP, Johnson LA, Ohyashiki K, Toyama K, et al. Fusion of the nucleoporin gene NUP98 to HOXA9 by the chromosome translocation t(7;11)(p15;p15) in human myeloid leukaemia. Nature genetics. 1996 Feb;12:154–158.

[360] Borrow J, Shearman AM, Stanton VP, Becher R, Collins T, Williams AJ, et al. The t(7;11)(p15;p15) translocation in acute myeloid leukaemia fuses the genes for nucleoporin NUP98 and class I homeoprotein HOXA9. Nature genetics. 1996 Feb;12:159–167.

[361] Kroon E, Thorsteinsdottir U, Mayotte N, Nakamura T, Sauvageau G. NUP98-HOXA9 expression in hemopoietic stem cells induces chronic and acute myeloid leukemias in mice. The EMBO journal. 2001 Feb;20:350–361.

[362] van Zutven LJCM, Onen E, Velthuizen SCJM, van Drunen E, von Bergh ARM, van den Heuvel-Eibrink MM, et al. Identification of NUP98 abnormalities in acute leukemia: JARID1A (12p13) as a new partner gene. Genes, chromosomes & cancer. 2006 May;45:437–446.

[363] Wang GG, Cai L, Pasillas MP, Kamps MP. NUP98-NSD1 links H3K36 methylation to Hox-A gene activation and leukaemogenesis. Nature cell biology. 2007 Jul;9:804–812.

[364] Wang GG, Song J, Wang Z, Dormann HL, Casadio F, Li H, et al. Haematopoietic malignancies caused by dysregulation of a chromatin-binding PHD finger. Nature. 2009 Jun;459:847–851.

[365] Franks TM, McCloskey A, Shokirev MN, Benner C, Rathore A, Hetzer MW. Nup98 recruits the Wdr82-Set1A/COMPASS complex to promoters to regulate H3K4 trimethylation in hematopoietic progenitor cells. Genes & development. 2017 Nov;31:2222–2234.

[366] Wong SHK, Goode DL, Iwasaki M, Wei MC, Kuo HP, Zhu L, et al. The H3K4-Methyl Epigenome Regulates Leukemia Stem Cell Oncogenic Potential. Cancer Cell. 2015 Aug;28(2):198–209. Available from: http://dx.doi.org/10.1016/j.ccell.2015.06.003.

[367] Vigorito E, Bell S, Hebeis BJ, Reynolds H, McAdam S, Emson PC, et al. Immunological function in mice lacking the Rac-related GTPase RhoG. Molecular and cellular biology. 2004 Jan;24:719–729.

[368] van Buul JD, Allingham MJ, Samson T, Meller J, Boulter E, García-Mata R, et al. RhoG regulates endothelial apical cup assembly downstream from ICAM1 engagement and is involved in leukocyte trans-endothelial migration. The Journal of cell biology. 2007 Sep;178:1279–1293.

[369] Tybulewicz VLJ. Vav-family proteins in T-cell signalling. Current opinion in immunology. 2005 Jun;17:267–274.

[370] Jackson BC, Ivanova IA, Dagnino L. An ELMO2-RhoG-ILK network modulates microtubule dynamics. Molecular biology of the cell. 2015 Jul;26:2712–2725.

[371] Wang JY, Yu P, Chen S, Xing H, Chen Y, Wang M, et al. Activation of Rac1 GTPase promotes leukemia cell chemotherapy resistance, quiescence and niche interaction. Molecular oncology. 2013 Oct;7:907–916.

[372] Nimmagadda SC, Frey S, Edelmann B, Hellmich C, Zaitseva L, König GM, et al. Bruton's tyrosine kinase and RAC1 promote cell survival in MLL-rearranged acute myeloid leukemia. Leukemia. 2018 Mar;32:846–849.

[373] Marschalek R. Systematic Classification of Mixed-Lineage Leukemia Fusion Partners Predicts Additional Cancer Pathways. Annals of laboratory medicine. 2016 Mar;36:85–100.

[374] Morris VA, Cummings CL, Korb B, Boaglio S, Oehler VG. Deregulated KLF4 Expression in Myeloid Leukemias Alters Cell Proliferation and Differentiation through MicroRNA and Gene Targets. Molecular and cellular biology. 2016 Feb;36:559–573.

[375] Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell. 2006 Aug;126:663–676.

[376] Liu Y, Cheng H, Gao S, Lu X, He F, Hu L, et al. Reprogramming of MLL-AF9 leukemia cells into pluripotent stem cells. Leukemia. 2014 May;28:1071–1080.

[377] Park SM, Cho H, Thornton AM, Barlowe TS, Chou T, Chhangawala S, et al. IKZF2 Drives Leukemia Stem Cell Self-Renewal and Inhibits Myeloid Differentiation. Cell stem cell. 2019 Jan;24:153–165.e7.

[378] Advani AS, Lim K, Gibson S, Shadman M, Jin T, Copelan E, et al. OCT-2 expression and OCT-2/BOB.1 co-expression predict prognosis in patients with newly diagnosed acute myeloid leukemia. Leukemia & lymphoma. 2010 Apr;51:606–612.

[379] Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. Cell. 2014 Jun;157:1262–1278.

[380] Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR-Cas9. Nature reviews Genetics. 2015 May;16:299–311.

[381] Dominguez AA, Lim WA, Qi LS. Beyond editing: repurposing CRISPR-Cas9 for precision genome regulation and interrogation. Nature reviews Molecular cell biology. 2016 Jan;17:5–15.

[382] Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, et al. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell. 2014 Oct;159:647–661.

[383] Sanjana NE, Shalem O, Zhang F. Improved vectors and genome-wide libraries for CRISPR screening. Nature methods. 2014 Aug;11:783–784.

[384] Bach C, Mueller D, Buhl S, Garcia-Cuellar M, Slany R. Alterations of the CxxC domain preclude oncogenic activation of mixed-lineage leukemia 2. Oncogene. 2009;28(6):815–823.

[385] Chen Y, Anastassiadis K, Kranz A, Stewart AF, Arndt K, Waskow C, et al. MLL2, Not MLL1, Plays a Major Role in Sustaining MLL-Rearranged Acute Myeloid Leukemia. Cancer Cell. 2017;31(6):755–770.

[386] Boeva V. Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells. Frontiers in genetics. 2016;7:24.

[387] Jayaram N, Usvyat D, R Martin AC. Evaluating tools for transcription factor binding site prediction. BMC bioinformatics. 2016 Nov;.

[388] Ruan S, Stormo GD. Inherent limitations of probabilistic models for protein-DNA binding specificity. PLoS computational biology. 2017 Jul;13:e1005638.

[389] Liu S, Zibetti C, Wan J, Wang G, Blackshaw S, Qian J. Assessing the model transferability for prediction of transcription factor binding sites based on chromatin accessibility. BMC bioinformatics. 2017 Jul;18:355.

[390] Yang J, Ramsey SA. A DNA shape-based regulatory score improves position-weight matrix-based recognition of transcription factor binding sites. Bioinformatics (Oxford, England). 2015 Nov;31:3445–3450.

[391] Isakova A, Groux R, Imbeault M, Rainer P, Alpern D, Dainese R, et al. SMiLE-seq identifies binding motifs of single and dimeric transcription factors. Nature methods. 2017 Mar;14:316–322.

[392] Le DD, Shimko TC, Aditham AK, Keys AM, Longwell SA, Orenstein Y, et al. Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. Proceedings of the National Academy of Sciences of the United States of America. 2018 Apr;115:E3702–E3711.

[393] Rastogi C, Rube HT, Kribelbauer JF, Crocker J, Loker RE, Martini GD, et al. Accurate and sensitive quantification of protein-DNA binding affinity. Proceedings of the National Academy of Sciences of the United States of America. 2018 Apr;115:E3692–E3701.

[394] Quang D, Xie X. FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. Methods (San Diego, Calif). 2019 Aug;166:40–47.

[395] Artinger EL, Mishra BP, Zaffuto KM, Li BE, Chung EKY, Moore AW, et al. An MLL-dependent network sustains hematopoiesis. Proceedings of the National Academy of Sciences of the United States of America. 2013 Jul;110:12000–12005.

[396] Hu D, Gao X, Cao K, Morgan MA, Mas G, Smith ER, et al. Not all H3K4 methylations are created equal: Mll2/COMPASS dependency in primordial germ cell specification. Molecular Cell. 2017;65(3):460–475.

[397] Viré E, Brenner C, Deplus R, Blanchon L, Fraga M, Didelot C, et al. The Polycomb group protein EZH2 directly controls DNA methylation. Nature. 2006 Feb;439:871–874.

[398] Stemmer M, Thumberger T, Del Sol Keyer M, Wittbrodt J, Mateo JL. CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. PloS one. 2015;10:e0124633.