Published in final edited form as:

Nat Genet. 2014 January ; 46(1): 17–23. doi:10.1038/ng.2836.

Large conserved domains of low DNA methylation maintained by Dnmt3a

Mira Jeong[1,#], **Deqiang Sun**[2,#], **Min Luo**[1,#], **Yun Huang**[3], **Grant A. Challen**[1,%], **Benjamin Rodriguez**[2], **Xiaotian Zhang**[1], **Lukas Chavez**[3], **Hui Wang**[4], **Rebecca Hannah**[5], **Sang-Bae Kim**[6], **Liubin Yang**[1], **Myunggon Ko**[3], **Rui Chen**[4], **Berthold Göttgens**[5], **Ju-Seog Lee**[6], **Preethi Gunaratne**[7], **Lucy A. Godley**[8], **Gretchen J. Darlington**[9], **Anjana Rao**[3], **Wei Li**[2,*,^], and **Margaret A. Goodell**[1,*,^]

[1]Stem Cells and Regenerative Medicine Center, Department of Pediatrics and Molecular & Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

[2]Division of Biostatistics, Dan L. Duncan Cancer Center and Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, Texas 77030, USA

[3]Division of Signaling and Gene Expression, La Jolla Institute for Allergy and Immunology, La Jolla, California 92037, USA

[4]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA

[5]Department of Hematology, Cambridge Institute for Medical Research and Wellcome Trust and MRC Cambridge Stem Cell Institute, Cambridge University, Hills Road, Cambridge, UK

[6]Department of Systems Biology, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX 77054, USA

[7]Department of Pathology, Baylor College of Medicine, and Department of Biology & Biochemistry, University of Houston, Houston, TX 77204, USA

[8]Department of Medicine, The University of Chicago, Chicago, Illinois 60637, USA

[9]Huffington Center for Aging, Baylor College of Medicine, Houston, Texas 77030, USA

Abstract

[*]Correspondence to: goodell@bcm.edu or WL1@bcm.edu.

[#]These authors contributed equally to this work

[^]These authors jointly directed this work

[%]Now at the Department of Internal Medicine, Washington University in St. Louis, St. Louis, Missouri 63110, USA.

Data access

All the data sets can be downloaded with GEO accession number (GSE49191) (www.ncbi.nlm.nih.gov/geo/), and UCSC browser tracks (mouse mm9) can be accessed with this hub: http://dldcc-web.brc.bcm.edu/lilab/benji/canyon.tracks.txt

AUTHOR CONTRIBUTIONS

Designed and performed experiments: MJ, ML, GAC, XZ, YH, MK, HW, LY, RC; Analyzed data: MJ, DS, ML, GAC, BR, LC, SK, RH, LAG, AR, GJD, WL, MAG. Wrote and edited the paper: MJ, DS, ML, GAC, BR, JL, BG, PG, LAG, GJD, AR, WL, MAG.

COMPETING FINALCIAL INTERESTS

The authors declare no competing financial interests.

Gains and losses in DNA methylation are prominent features of mammalian cell types. To gain insight into mechanisms that could promote shifts in DNA methylation and contribute to cell fate changes, including malignant transformation, we performed genome-wide mapping of 5-methylcytosine and 5-hydroxymethylcytosine in purified murine hematopoietic stem cells. We discovered extended regions of low methylation (Canyons) that span conserved domains frequently containing transcription factors and are distinct from CpG islands and shores. The genes in about half of these methylation Canyons are coated with repressive histone marks while the remainder are covered by activating histone marks and are highly expressed in HSCs. Canyon borders are demarked by 5-hydroxymethylcytosine and become eroded in the absence of DNA methyltransferase 3a (Dnmt3a). Genes dysregulated in human leukemias are enriched for Canyon-associated genes. The novel epigenetic landscape we describe may provide a mechanism for the regulation of hematopoiesis and may contribute to leukemia development.

The majority of cytosines adjacent to guanines (CpGs) in the mammalian genome are methylated (5mC) except in gene regulatory regions where they are often clustered and unmethylated (CpG islands, CGI) [1]. Although regions of low CpG methylation are considered generally permissive for gene expression when present in promoter regions, we still understand only poorly how DNA methylation patterns vary among normal cell types, how they are added and erased, and how they influence gene expression. While CGIs tend to exhibit low levels of methylation across many cell types, the greatest variation in DNA methylation levels across different cell types is thought to occur primarily in regions adjacent to CGIs, termed “shores” that are also hotspots for hyper- and hypo-methylation in malignant cells[2]. However, most of our understanding of changes in DNA methylation patterns comes from limited analysis of cell lines, tissues of heterogeneous composition, or cancer cells whose lineal relationships are not always well understood. Moreover, identification of recurrent leukemia-associated mutations in genes encoding regulators of DNA methylation such as DNMT3A and TET2 [3–6] have underscored the critical importance of DNA methylation in maintenance of normal physiology. To gain insight into how DNA methylation exerts this central role, we sought to determine the genome-wide pattern of DNA methylation in the normal precursors of leukemia cells: the hematopoietic stem cell (HSC), and investigate the factors that affect alterations in DNA methylation and gene expression.

RESULTS

The murine HSC DNA methylome

We performed whole genome bisulfite sequencing (WGBS) on purified murine HSCs (side population (SP) cells that were also lineage-marker-negative, c-Kit+ Sca-1+ and CD150+; please see methods) with two biological replicates achieving a total of 1,121M reads, of which 80.2 % were successfully aligned to either strand of the reference genome (mm9), resulting in a combined average of 40X coverage (Supplementary Table 1). There were two replicates and the data were highly reproducible with a correlation coefficient of more than 0.99 between methylation ratios genome-wide for both phenotypes. In general, the HSC methylome was similar to that of other mammalian cells[7,8]. DNA methylation was low in CpG islands (CGI) and promoters, and higher in gene bodies and repetitive elements

(Supplementary Fig. 1). In addition, non-CpG methylation was infrequent (less than 1% CpH methylation), consistent with other non-ES cell types[9].

Identification of large under-methylated Canyons with unique genomic features

Previous WGBS studies demonstrated that hypomethylated regions are enriched for functional regulatory elements such as promoters and enhancers[8,10]. Here, we used a Hidden Markov Model to identify under-methylated regions (UMRs) with average proportion of methylation ≤ 10% (Supplementary Table 2) and required at least 5 CpGs per kb to satisfy the permutation-based FDR 5%. Using these criteria, there are 32,325 UMRs in mouse HSC methylome. Most UMRs are associated with promoters or gene bodies and only 8.3% showed intergenic localization. By inspecting the UMR size distribution, we observed that a small portion were exceptionally large, with some of them extending over 25 kb, such as the UMR associated with the *Pax6* gene (Fig. 1a), representing an expanse of unmethylated DNA that is considerably larger than that previously reported. In the genome landscape, these large methylation-depleted regions appear as “canyons” cut into a plateau of high methylation, usually sequestering a single gene.

In order to determine whether these large UMRs represented some unique genomic feature, we required them to be at least 3.5 Kb in length (>10 times larger than the typical CGI of ~300 bp[11]; and see methods); this revealed 1,104 methylation “Canyons” representing 3.4% of all UMRs (Supplementary Fig. 2, Supplementary Table 2). To compare these with typical UMRs, we established a control group of 13,579 UMRs (cUMRs) that were longer than 1kb but smaller than 3.5 kb. This control group eliminates the smallest UMRs that tend to be transcription factor binding sites. To gain insights into the biological function of these Canyons, we performed gene ontology enrichment analysis with Canyon-associated genes and cUMR-associated genes. Canyon-associated genes showed a striking pattern of enrichment for genes involved in transcriptional regulation (318 genes, P=6.2 × 10^{-123}), as well as genes containing a homeobox domain (111 genes, P=3.9 × 10^{-85}) (Fig. 1b, Supplementary Fig. 3a, Supplementary Table 3), comprising one of the most ancient gene families involved in embryonic development of bilaterians. In contrast, the genes associated with cUMRs all give non-significant p-values on these 4 GO terms. Among the largest 20 Canyons, 15 harbor homeobox-containing genes (Supplemental Fig. 3b). These Canyons typically extend well outside of the immediate coding regions of these genes (Supplementary Fig. 3c–e). As a group, these Canyons are particularly highly conserved (Supplementary Fig. 4a) and are depleted for transposable elements and repeats (Supplementary Fig. 4b–c).

The vicinity of developmental genes has previously been noted to be depleted of recent transposable element insertions;[12] the pattern of repeat insertion may contribute to attracting DNA methylation outside Canyons. Interestingly, some of the homeobox-containing orthologs in *D. melanogaster*, which lacks DNA methylation, also are associated with higher promoter CpG content[13] and are also resistant to transposable element insertion[14].

We noted that the conservation and pattern of gene ontology enrichment for Canyon genes was similar to that described for a group of genes considered to be targets of highly conserved non-coding elements within large (~1 megabase) genomic regulatory blocks

(GRBs)[15,16]. To test this systematically, we examined the relationship between UMR size and GRB target genes. We plotted overlap in membership with GRB genes and three control gene groups against membership in UMR gene groups established by different UMR length cutoffs (Fig. 1c). We found that the group of UMRs that are ≥3.5 kb overlap with 67% of the GRB targets while the remaining 31,221 UMRs ≤3.5 kb overlap with only 27% of GRB targets (P=2×E-16). This analysis suggests that methylation Canyons are key elements of ancient gene regulatory domains.

To better understand these Canyons, we compared them with other genomic features associated with low levels of DNA methylation. While CGI are present in most Canyons, 10% do not contain a classically-defined CGI [11], and 53% contain a single CGI and are only covered by CGI at a median of 26%; therefore, the presence of CGI cannot alone explain these methylation lacunae (Fig. 1d). CGI shores, which flank 2kb on either side of CGI, have been shown to exhibit the greatest methylation variation across cell types[2]. Because most Canyons contain one or more CGI, they will also harbor associated shores.

Recently it was reported that there are large genomic domains called Super-enhancers occupied by master transcription factors and the mediator complex[17]. While these have not been defined in HSCs by mediator, we reasoned that sites in which multiple TFs bind across several hematopoietic cell types would approximate such regions. To examine their relationship with Canyons, we compared Canyons with TF binding sites as identified from more than 150 ChIP-seq data sets across a variety of blood lineages (>10)[18]. Interestingly, we found that TF binding peaks for 10 HSC pluripotency TFs are significantly enriched not only in small cUMR regions but also across the entirety of Canyons compared with their surrounding regions (Fig. 1e and Supplementary Fig. 5a–c).

Methylation Canyons are conserved among cell types and species

To determine whether Canyons are stable or cell-type-variable features, we identified Canyons in ESC methylome data [8] using the same criteria as for HSC. Of 839 ESC Canyons (Supplementary Table 4), 82% (688) were largely shared between both cell types, although there were variations in their edges, size, and average methylation levels (Supplementary Fig. 6a–e). Similarly, many Canyons identified in murine HSC could be identified in human hematopoietic progenitors and differentiated progeny (Supplementary Fig. 6f)[10] and non-hematopoietic cells, with minimal cell-type variation (Fig. 1f, Supplementary Fig. 6g). We found that 72%~80% of Canyons defined in mouse ES cells overlapped with Canyons in methylome data of a variety of human cell types (Supplementary Fig. 6h). These data establish that methylation Canyons are a distinct genomic feature that is stable, albeit with subtle differences, across cell-types and species. While most contain CGIs and shores, their methylation levels are generally exceedingly low and minimally variable, in contradistinction to the majority of shores found associated with CGI excluded from Canyons.

Expression of Canyon genes is regulated by histone modifications

Low DNA methylation is usually associated with active gene expression. However, many Canyon-associated genes are developmental regulators that are not known to play roles across many cell and tissue types thus we examined their regulatory features in more detail in the hematopoietic system. RNA-seq data indicated that among the twenty largest Canyons, only two harbored highly expressed genes: *Hoxa9* and *Meis1*, which encode transcription factors critical for hematopoiesis and frequently deregulated in leukemia (Supplementary Fig. 3a). To examine whether histone modifications could account for the lack of expression from other Canyon genes, we investigated their activating H3K4me3 and repressive H3K27me3 histone marks by ChIP-seq. While most cUMRs were associated with high H3K4me3 and low H3K27me3 marks, the Canyons showed a distinct bi-modal distribution, with around half exhibiting high H3K4me3 and half high H3K27me3 marks (Fig. 2a–b)

Among all the murine HSC Canyons, 6% were only bound by H3K27me3 and 45.9% exhibited both H3K27me3 and H3K4me3 binding, similar to the so-called “bivalent” domains found in embryonic stem cells (e.g. *Gata6*, Fig. 2c)[19]. The H3K27me3 often covered the entire length of the Canyon, such that the Canyon edges were aligned with the ends of the H3K27 me3 peak, as in the *Uncx* gene (Fig. 2d). Similarly, the remaining H3K27me3-negative Canyons were heavily coated by H3K4me3, e.g. the *Meis1* gene (Fig. 2e). H3K27me3 was the defining feature for expression, as the H3K4me3-only Canyon genes were highly expressed, while the H3K27me3-associated genes showed low or no expression regardless of their H3K4me3 association and the median expression level of H3K4me3-only Canyon genes was higher than comparable cUMR (Fig. 2f). While in HSCs we cannot determine whether individual cells harbor both activating and repressive histone marks at the same allele, these data are consistent with a special epigenetic status of a certain subset of developmentally important genes, in which they exhibit activation-associated DNA methylation lacuna along with the repressive H3K27me3 mark, as well as (at most loci) some association with the activating H3K4me3 mark. In ES cells these “bivalent” loci have been proposed to represent a poised state in which these loci will be expressed during differentiation. Their putative presence in HSCs suggests instead that they reflect either a privileged epigenetic status or perhaps indicate differentiation history rather than future potential.

Methylation Canyons partially overlap with but are distinct from other low-methylation regions

Most studies of DNA methylation have largely focused on CpG islands, defined as being more than 300bp in length and having over 50% CG composition, where they are unmethylated and generally associated with promoters. Recent genome-wide approaches have revealed additional regions with important methylation alterations in cancer and cell fate decisions, such as CGI shores[2], partially methylated domains (PMD)[7], low methylated regions (LMR)[20], and long-range epigenetic activation (LREA)/suppression (LRES) regions[21,22]. Here, we established the presence of a distinct hypomethylated feature that is highly conserved and stable across cell types and species. While these methylation Canyons share many features of smaller undermethylated regions, they represent only 3.4% of all UMRs and are distinct in their very low methylation level, their enrichment for homeobox-containing genes, their overlap with the GRB target genes, their stability between cell types, and their bimodal distribution of H3K27me3 that indicates a distinct mode of gene expression regulation (Supplementary Table 5).

Methylation Canyons are maintained by Dnmt3a

Because *DNMT3A* is mutated in a high frequency of human leukemias[23], we examined the impact of loss of Dnmt3a on Canyon size. We compared all UMRs in HSCs conditionally inactivated for *Dnmt3a* (KO) to wild-type (WT) HSCs. Upon knockout of *Dnmt3a*, the edges of the cUMRs and Canyons are hotspots of differential methylation while regions inside of cUMRs and Canyon are relatively resistant (Supplementary Fig. 7a). Thirty percent of all differentially methylated regions (DMRs) in the *Dnmt3a* KO were located at the edges of UMRs. This focused methylation loss at the edges of UMRs suggests that Dnmt3a normally acts to maintain methylation at their boundaries (Fig. 3a and Supplementary Fig. 7a). On 44% of Canyons, the edges were eroded such that they increased in size, and 31% of Canyons experienced hypermethylation at the edges, such they decreased in size (25% experienced no significant change). The methylation loss in *Dnmt3a* KO HSCs led to the addition of 861 new Canyons for a total of 1787 Canyons (Fig. 3b, Supplementary Table 6). Methylation in some regions that featured a cluster of Canyons in WT HSCs was decimated such that Canyons merged to become groups of larger Canyons (“Grand Canyons”), as exemplified by the *HoxB* region, in which the enlarged Canyon covers more than 50 kb, interrupted by short stretches of higher methylation (Supplementary Fig. 7b).

The expansion and contraction of different Canyons in absence of Dnmt3a is reminiscent of the concomitant hyper- and hypo-methylation that is observed in many malignant cells; thus, we considered whether other epigenetic mechanisms influenced Canyon behavior. We first examined the histone mark distribution on expanding vs. contracting Canyons. For the WT-defined Canyons, those marked with H3K4me3 only were most likely to expand after Dnmt3a KO. In contrast, the canyons marked only with H3K27me3 or with both marks were more likely to contract (Fig. 3c). This suggests Dnmt3a specifically is acting to restrain Canyon size where active histone marks (and active transcription) are already present.

Methylation Canyon borders are demarked by 5-hydroxymethylcytosine

We next considered whether Canyon edge erosion was attributable to an active process. The Tet protein family may promote demethylation, as hydroxy-methylated cytosine is not recognized by Dnmt1, leading to replacement of 5hmC with unmethylated cytosine during DNA replication [24,25]. WGBS cannot distinguish between 5mC and 5hmC, so we determined the genome-wide distribution of 5hmC in WT and Dnmt3a KO HSCs using the cytosine-5-methylenesulphonate (CMS)-Seq method [26] in which sodium bisulfite treatment converts 5hmC to CMS; CMS-containing DNA fragments are then immunoprecipitated using a CMS-specific antiserum (Supplementary Table 7). Several sites of CMS signal were validated using oxBS-sequencing[27], based on quantitative sequencing of 5mC and 5hmC at single base resolution (Supplementary Table 8, Supplementary Fig. 8a, 8b). Strikingly, 5hmC peaks were enriched specifically at the borders of both cUMRs and Canyons (Supplementary Fig. 8c). In particular, expanding Canyons, typically associated with highest H3K4me3 marking, were highly enriched at the edges for the 5hmC signal (Fig. 4a, 4b). In contrast, contracting Canyons, more likely to be associated with H3K27me3, were depleted of 5hmC (Fig. 4a, 4c). An example of an expanding Canyon is the HSC master regulator *Gata2*, which shows 5hmC peaks at the Canyon boundary in WT HSCs, and methylation edge erosion in the *Dnmt3a* KO (Fig. 4d). Where the methylation signal is completely depleted in the *Dnmt3a* KO, 5hmC peaks disappear altogether, consistent with loss of the 5mC substrate for hydroxylation. Where methylation is merely reduced, the 5hmC signal tended to increase (Supplementary Fig. 9a–d) suggesting unimpeded access to the DNA by the Tet proteins. We may expect additional divisions of the *Dnmt3a* KO HSCs would result in elimination of methylation at these sites, possibly contributing to further decline of their differentiation potential [28]. It is worth noting that the *DNMT3A* somatic mutations found in AML patients are distinct from the *Dnmt3a* null allele used here, with many patients being heterozygous for R882, a specific catalytic domain point mutation[3] and others being compound heterozygous for likely inactivating mutations[29], so the impact of patient *DNMT3A* mutations on Canyon edges may be different. All three Tet family proteins are expressed in murine HSCs, thus we cannot determine which contribute to establishing / maintaining the 5hmC signal. Furthermore, direct action on the Canyon edges by Dnmt3a and specific Tet proteins ultimately needs to be established with biochemical methods.

Methylation Canyon gene expression is associated with cancer

Aberrant hyper-methylation in transformed cells has been thought to contribute to malignancy development [30], and both hyper- and hypo-methylation is associated with transformed cells. Thus, we tested whether Canyon-associated genes were likely to be associated with hematologic malignancy development. We used Oncomine to assess whether Canyon genes expressed in WT HSCs were associated with the aberrant expression signatures of human Leukemias. These Canyon genes were highly enriched in seven signatures of genes over-expressed in Leukemia patients compared to normal bone marrow; in contrast, four sets of control genes were not similarly enriched (Fig. 5a, Supplementary Table 9). Further, we used TCGA data to test whether Canyon gene expression changes were associated with *DNMT3A* mutation in AML patients. Remarkably, we found that expressed canyon genes are significantly enriched for differentially expressed genes between patients with and without *DNMT3A* mutation (p value<0.05) (Fig. 5b, Supplementary Table 9). Overall, 76 expressed canyon genes, including multiple HOX genes, are significantly changed in patients with *DNMT3A* mutation (p=0.0031) (Supplementary Table 9). Notably, the previous gene expression comparison in whole transcriptome level did not identify any expression cluster associated with *DNMT3A* mutation[3], but we identified two strong clusters from unsupervised clustering with 80% of Dnmt3a mutant patients enriched into cluster A (Supplementary Fig. 10). The expressed canyon genes identified here may be used as a unique gene expression signature to define the *DNMT3A* mutation status in patients. We further checked Canyon genes expressions in various other cancer types by using data from cancer cell line encyclopedia (CCDE; a compilation of gene expression data from 947 human cancer cell lines). Canyons expressed in HSC are highly expressed or depleted in hematologic cancer cell lines, whereas unexpressed canyons showed high expression in other cancer cell lines, which may reflect the original tissue-specificity of canyon expressions which regulated by histone modification (Supplementary Fig. 11).

DISCUSSION

Here we have demonstrated the existence of very large methylation lacunae associated with highly conserved developmentally important genes. Expression of genes in many methylation Canyons is restrained by broad H3K27me3-marked polycomb-regulated zones, whereas active Canyons exhibit high H3K4me3 Trithorax-associated marking. Similar features harboring developmental regulators have been noted in other species [31] and recently in ES cells, where they were termed DNA methylation valleys (DMVs) [32]. DMVs, defined by slightly different methylation level and size criteria, include 1220 genomic conserved loci enriched for developmental regulators, which were also marked by either H3K4me3 or H3K27me3.

The active HSC canyons, containing genes involved in hematopoiesis and frequently dysregulated in leukemias, are particularly susceptible to DNA methylation loss. This suggests a model in which Tet proteins and Dnmt3a act concomitantly on Canyon borders (Supplementary Fig. 12), opposing each other in alternately effacing and restoring methylation at the edges, particularly at sites of active chromatin marks. The insight that Tets and Dnmt3a compete to maintain the *status quo* at the same loci in HSCs enables multiple scenarios to be envisioned in which action of one protein or the other is reduced, either due to gene expression attenuation or by mutation, leading to subsequent consequences on methylation, gene expression, and developmental potential.

The observation that quiescent Canyons do not expand with Dnmt3a loss, and often shrink, suggests that Dnmt3b or other mechanisms drive hypermethylation specifically associated with H3K27me3 marks. The genes in these Canyons are generally not associated with hematologic malignancies; these Canyons may be largely inert in this lineage, despite significant epigenetic perturbation in the transformed state.

Mutations in *DNMT3A* and *TET2* have been linked to a similar spectrum of hematologic malignancies [3–5,33]. Although the proteins appear to oppose each other biochemically, genetically, their mutations have a similar impact in impeding differentiation and promoting transformation. While the precise mechanisms through which this occurs are still unclear, the action of Dnmt3a and Tets at the same genomic sites may suggest that imbalance in either disrupts the broader regulatory mechanisms acting at these loci. The reported poor correlation between methylation changes and gene expression changes in both mouse models [28] and human patient samples [3] may reflect the complex regulation at these loci, indicating the need to take multiple epigenetic factors into account as we seek to understand the pathogenesis of these malignancies.

Online Methods

Hematopoietic Stem Cell Purification and Flow Cytometry

For WT HSCs, whole bone marrow cells were isolated from femurs, tibias, pelvis and humerus of 12 month-old male C57Bl/6 mice. 10 mice were used to purify HSCs; biological replicates were performed with two separate pools of HCSs from different donors. *Dnmt3a*-KO HSCs were purified from mice at the tertiary stage of serial transplantation, because at this point, the phenotype resulting from loss of *Dnmt3a* manifests most significantly[28]. 18-weeks after the tertiary transplants, donor cell derived (CD45.2+) HSCs were purified from four to eight transplanted mice per biological replicate. This timing allowed aged-matched comparison to 12-month-old wild-type HSCs.

HSCs from both WT and *Dnmt3a* KO mice were purified using the side population (SP)[34] strategy of Hoechst staining in combination with surface markers[35]. Briefly, whole bone marrow cells were resuspended in staining media at 10^6 cells/mL and incubated with 5 mg/mL Hoechst 33342 (Sigma) for 90 minutes at 37°C. For antibody staining, cells were suspended at a concentration of 10^8 cells/mL and incubated in 4°C for 15 minutes with the desired antibodies. Magnetic enrichment was performed with c-Kit-biotin antibody (eBioscience, San Diego, CA) and anti-biotin microbeads (Miltenyi Biotec, Auburn, CA) or anti-mouse CD117 microbeads (Miltenyi Biotec, Germany) on an AutoMACS (Miltenyi Biotec, Germany). Post-enrichment, the positive cell fraction was labeled with antibodies to identify HSCs (SP^+ $Lineage^-$ $Sca\text{-}1^+$ $c\text{-}Kit^+$ $CD150^+$). All antibodies were obtained from BD Biosciences (San Jose, CA) or eBioscience (San Diego, CA) and used at 1:100 dilutions. Cell sorting was performed on a MoFlo cell sorter (Dako North America, Carpinteria, CA) or Aria II (BD Biosciences, San Jose, CA) and analysis performed on a LSRII (BD Biosciences, San Jose, CA). All animal work was performed with approval from the Baylor College of Medicine Institutional Animal Care and Use Committee.

Whole-genome bisulfite sequencing (WGBS)

For WGBS library construction, 300 ng genomic DNA was isolated from HSCs and fragmented using a Covaris sonication system (Covaris S2). Following fragmentation, libraries were constructed using the Illumina TruSeq DNA sample preparation kit. After ligation, libraries were bisulfite-treated using the EpiTect Bisulfite Kit (Qiagen, Valencia, CA). Ligation efficiency tested by PCR using TrueSeq primers and Pfu TurboCx hotstart DNA polymerase (Stratagene). After determining the optimized PCR cycle number for each sample, a large scale PCR reaction (100ul) was performed as described previously[36]. PCR products were sequenced with Illumina HiSeq sequencing systems.

Anti-CMS technique[37] for detection of 5-hydroxymethylcytosine

For CMS precipitation, 1.5 μg of genomic DNA fragments were ligated with methylated adaptors and treated with sodium bisulfite (Qiagen). The DNA was then denatured for 10 min at 95 °C (0.4 M NaOH, 10 mM EDTA), neutralized by addition of cold 2 M ammonium acetate pH 7.0, incubated with anti-CMS antiserum in 1× immunoprecipitation buffer (10 mM sodium phosphate pH 7.0, 140 mM NaCl, 0.05% Triton X-100) for 2 h at 4 °C, and then precipitated with Protein G beads. Precipitated DNA was eluted with Proteinase K, purified by phenol-chloroform extraction, and amplified by 8 cycles PCR using Pfu TurboCx hotstart DNA polymerase (Stratagene). DNA sequencing was carried out using Illumina/Solexa Genome Analyzer II and HiSeq sequencing systems.

OxBS sequencing

Genomic DNA was further purified by ethanol precipitation and micro Bio-Spin 6 column (Bio-Rad). 250 ng purified genomic DNA was denatured in at 24 μl of 0.05 M NaOH at

37°C for 30 min, and then snap cooled on ice for 5 min. Next, 1 µl of $KRuO_4$ (Sigma) (15 mM in 0.05 M NaOH) was added to denatured gnomic DNA on ice for 1 hour, with occasional vortexing. The mixture was purified with micro Bio-Spin 6 column. The non-oxidized and oxidized genomic DNAs were treated with MethylCode bisulfite conversion kit (Invitrogen). Loci-specific PCRs were performed using PyroMark PCR kit (Qiagen). Amplicons were pooled together and barcoded libraries were prepared by TruSeq library preparation kit (Illumina). Amplicon sequencing was performed on MiSeq (Illumina).

Computational analysis oxBS sequencing data

Bisulfite and oxidative bisulfite sequencing data were mapped against mm9 using the Bismark software (PMID 21493656) v0.6.4 (-q -n 2 --chunkmbs 1028 bowtie-0.12.7). Subsequently, number of reads containing converted and the number of reads containing unconverted cytosines at covered cytosines were counted based on Bismark's mapping results using custom scripts. For CpGs covered by at least 100 reads in both, the BS and oxBS sample, the percentage of hydroxymethylation has been calculated by subtracting the observed methylation in ox-bisulfite from the observed methylation in bisulfite.

RNA-sequencing (RNA-seq)

~70,000 HSCs were sorted into Trizol from the pools of each age group. RNA was isolated with the RNeasy Micro column (Qiagen, Valencia, CA). Paired end libraries were generated by using Illumina TruSeq RNA sample preparation kit. Illumina HiSeq was used for sequencing with a paired-end sequencing length of 100bp.

ChIP-sequencing (ChIP-seq)

Chromatin Immunoprecipitation (ChIP) was performed as described previously[38]. Briefly, 20,000~50,000 HSCs ($SP^{KLS}CD150^+$) were sorted and crosslinked with 1% formaldehyde at room temperature (RT) for 10 min, and the reaction was stopped by 0.125M glycine at RT for 5 min. Then the cells were washed once with ice cold PBS containing protease inhibitor cocktail (PIC; Roche) and the cell pellet was stored at −80°C. Cross-linked cells were thawed on ice and lysed in 50 µl Lysis buffer (10 mM Tris pH 7.5, 1mM EDTA, 1% SDS), then diluted with 150 µl of PBS/PIC, and sonicated to 200–500 bp fragments (Bioruptor, Diagenode). The sonicated chromatin was centrifuged at 4°C for 5 min at 13,000rpm to remove precipitated SDS. 180 µl was then transferred to a new 0.5 ml collection tube, and 180 µl of 2X RIPA buffer (20 mM Tris pH 7.5, 2 mM EDTA, 2%Triton X-100, 0.2% SDS, 0.2% sodium deoxycholate, 200 mM NaCl/PIC) was added to recovered supernatants. A 1 /10 volume (36 µl) was removed for input control. ChIP-qualified antibodies (0.1 µg H3K4me3 Millipore 07-473, 0.3 µg H3K27me3 Millipore 07-449) were added to the sonicated chromatin and incubated at 4°C overnight. Following this, 10 µl of protein A magnetic beads (Dynal, Invitrogen) previously washed in RIPA buffer were added and incubated for an additional 2 hours at 4°C. The bead: protein complexes were washed three times with RIPA buffer and twice with TE (10 mM Tris pH 8.0/1 mM EDTA) buffer. Following transfer into new 1.5 ml collection tube, genomic DNA was eluted for 2 hours at 68 °C in 100 µl Complete Elution Buffer (20 mM Tris pH 7.5, 5 mM EDTA, 50 mM NaCl, 1% SDS, 50 µg/ml proteinase K), and combined with a second elution of 100 µl Elution Buffer (20 mM Tris pH 7.5, 5 mM EDTA, 50 mM NaCl) for 10 min at 68 °C. ChIPed DNA was purified by MinElute Purification Kit (Qiagen) and eluted in 12 μl elution buffer. ChIPed DNA were successfully made library using ThruPLEX-FD preparation kit without extra amplification (Rubicon, Ann Arbor, MI). Sequencing was performed according to the manufacturer’s protocol on a HiSeq 2000 (Illumina). Sequenced reads were mapped to the mm9 mouse genome and peaks were identified by model-based analysis of ChIP-seq data (MACS).

Analysis of Whole Genome Bisulfite Sequencing (WGBS) Data

The WGBS data analyses were based on BSMAP[39] and a newly developed program MOABS: MOdel based Analysis of Bisulfite Sequencing (Sun *et al.* http://code.google.com/p/moabs/, manuscript in preparation). We have used four modules of MOABS, mMap, mCall, mOne and mComp from this software. MOABS seamlessly integrates alignment, methylation ratio calling, and identification of hypomethylation for one sample and differential methylation for multiple samples, and other downstream analysis.

Reads Mapping

BSMAP[39] was used to align the paired-end bisulfite treated reads to the mouse genome mm9. The adaptor and low quality sequences were automatically trimmed by BSMAP. For each read, the mapping location was determined to be the location with the fewest mismatches. If a read can be mapped to multiple locations with the same fewest mismatches, this read is determined as a multi-mapped read and its mapping location was randomly selected from all mapping locations.

Quality control and methylation ratio calling

BSeQC[40] was used to remove technical biases in WGBS data. First, we removed clonal reads with identical sequences resulting from possible over-amplification during sample preparation. These clonal reads were mapped to exactly the same position on the genome, and can be determined based on their extremely high coverage relative to the mean coverage across the genome based on a Poisson P value cutoff of 1×10^{-5}. As a result, at most 2 reads mapped to the same location were kept for the downstream analysis. Second, during adapter ligation in bisulfite library preparation, the overhangs of DNA fragments are end-repaired using unmethylated cytosines. This end repair procedure may introduce artifacts if the repaired bases contain methylated cytosines. We modeled the overhang size of DNA fragment and determined that trimming 3 bases (e.g. overhang size) from the repaired end was sufficient to eliminate nearly all artifact introduced by end-repair. Third, the overlapping segment of two read mates derived from the same DNA fragment was only processed once to prevent over-counting of the same DNA. Finally, methylation ratio of each CpG was measured as the proportion of unconverted CpGs in all mapped reads, including both strands.

Differentially methylated regions (DMRs)

We used a first order Hidden Markov Model (HMM) to determine differentially methylated regions (DMRs). For a two-sample comparison $p_2 - p_1$, the state of the i^{th} CpG in the genome is denoted as S_i where S_i can take 3 hidden states:

S_0: hypo-methylation state, if $p_2 - p_1 < -v_0$

S_1: no difference state, if $|p_2 - p_1| < v_0$

S_2: hyper-methylation state, if $p_2 - p_1 > v_0$

where v_0 is a preset threshold of methylation difference between two samples. We modeled the neighbor correlation by first order Markov chain: $\Pr(S_i) = \Pr(S_i|S_{i-1})$, where S_i is directly influenced by the state of previous CpG S_{i-1}.

For each CpG in the genome, we observed in total 4 numbers from 2 samples: $x = (n_1, k_1, n_2, k_2)$, where n is the number of mapped reads and k is the number of unconverted CpG in all mapped reads. Given the observation from all CpGs, we want to find the HMM model that maximizes the probability of the observation. The HMM is characterized by initial state π_0, transition probability matrix $A = Pr(S_i|S_{i-1})$ and emission probability matrix $B = Pr(x_i|S_i)$. The initial state π_0 can be assigned as S_1. By assuming a cytosine is in one of the three states, the emission probability for the i^{th} CpG as $x = (n_1, k_1, n_2, k_2)$, when the state of the cytosine is S_i, can be derived as

$$\Pr(n_1, k_1, n_2, k_2|s_i) = \frac{\iint_{s_i} dp_2 dp_1 f(k_1; n_1, p_1) f(k_2; n_2, p_2)}{\int_0^1 f(k_1; n_1, p_1) dp_1 \int_0^1 f(k_2; n_2, p_2) dp_2}$$

The transition probability matrix can be trained using the forward-backward algorithm. In the training process, the initial state, and the emission probability matrix are fixed while the state transition probability is the only model variable. Since the training is computationally intensive, MOABS chooses only a subset of CpGs for the training, such as the first one million CpGs in chromosome 19 or CpGs provided by the user. After the change of likelihood of the model is smaller than a given threshold or the max number of iterations is reached, the optimal hidden state for each CpG is obtained. Consecutive CpGs with the same hypo- or hyper- methylation state were merged as DMRs.

Under-methylated regions (UMRs)

Similar to DMR detection, we used a two-state first order Hidden Markov Model (HMM) to detect highly methylated and lowly methylated regions from a single sample. Only locations with coverage more than 10 reads were considered to increase the detection accuracy. Consecutive CpGs with the same hidden “low methylation” state were merged to form a low-methylation region (LMR). We also performed a random shuffle of all the CpGs in the genome, followed by the same procedure for LMR detection. The resulting NULL distribution indicates the number of CpGs required for LMR detection. With false discovery rate (FDR) at 5%, each LMR will include at least 4 CpGs for WT HSC or at least 5 CpGs for Dnmt3a-Knockout HSC. The UMRs are a subset of LMRs with mean methylation ratio less than 10%. Several highly methylated CpGs may separate two neighboring UMRs. We merged two such UMRs into a single UMR if the mean methylation ratio of the newly merged UMR is still less than 10%. UMRs less than 1kb long not used in this manuscript. UMRs greater than or equal to 3.5kb long were defined as "Canyon". UMRs greater than or equal to 1kb but less than 3.5kb are used as control UMRs (cUMRs) to compare with Canyons to show that Canyons are very unique. See (http://code.google.com/p/moabs/).

Analysis of 5hmc CMS pull down and histone modification ChIP-seq data

The 5-hydroxy-methylation CMS samples were sequenced at paired-end 100bp long. The reads were mapped to the mouse genome mm9 using BSMAP[39] by allowing at most 4 mismatches. Only uniquely mapped reads were used for MACS[41] peak calling at p-value cutoff E-5. Peaks are regions with enrichment in CMS pull down sample compared to control sample. The control is sonicated sample followed by bisulfite conversion but without CMS pull down.

The common peaks are those that overlap between wild-type and knock-out samples, while the sample specific peaks are those that do not overlap. To quantitatively detect the difference between two samples, all peaks from both samples were merged to form a new set of synthetic peaks, based on which a Poisson test was performed to detect if one sample has more reads than the other sample in each synthetic peak. Before the test, the read number is normalized to 10 million for every sample.

The same pipeline above was used to analyze histone modification ChIP-seq data, with a few exceptions. The histone modification ChIP-Seq reads were mapped to mouse genome mm9 using SOAP2[42] by allowing at most 2 mismatches for 50bp long short reads and at most 4 mismatches for 100bp long short reads. Only uniquely mapped reads are kept. To remove PCR resulted duplicate reads, at most 2 duplicate reads are allowed for each biological replicate. The number 2 is based on Poisson P-value cutoff of 1×10^{-5} determined by the total number of reads with respect to the theoretical mean coverage across the genome. The uniquely mapped and duplicate removed reads from each biological replicate are fed as treatment file into program MACS, to find the enriched regions, "peaks". The H3K4me3 peaks are called by MACS with default parameters except pvalue set at 1e-8. The H3K27me3 peaks are called by SICER with parameters "window size 200 fragment size 200 gap size 600 and FDR 1E-8". The peaks from all biological replicates of a specific sample are merged to form the final set of peaks for this specific sample.

Analysis of RNA-Seq data

Paired-end 100bp reads were sequenced for RNA-seq. The last 20 bases were trimmed due to average low quality. The alignment was performed by RUM [43], which first mapped reads to the genome and transcriptome by Bowtie, and then used blat to re-map those initially unmapped reads to the genome. The information from the two rounds of mappings was merged. The multiply mapped reads were discarded. The gene annotations used for transcriptome alignment include refSeq, UCSC knownGene and ensemble gene models. The gene expression, FPKM value, was calculated by counting the reads matching the exons of each gene. Differential expression was performed using edgeR [44].

UMR dynamics in size and methylation ratio

We defined the UMR dynamics in size including expanded, shrunk and unchanged between wild-type and knock-out samples using the following criterion: If one edge of a wild-type UMR moves outward, or inward in the knockout sample, for more than 200 bases, this edge is classified as “expanded” or “shrunk”, respectively. If the change is less than 200 bases, the edge is classified as “unchanged”. Furthermore, if the wild-type UMR disappears in knock-out sample, both edges of the UMR are classified as “shrunk”; whereas both edges of an emerging new UMR in knock-out sample are classified as “expanded”.

We merged all UMRs in both wild-type and knock-out samples into 19,569 synthetic UMRs and measured the contribution of each sample to the length of each synthetic UMR. The sample specific contribution to a given synthetic UMR is defined as length of sample specific UMR divided by length of synthetic UMR. The contribution from knock-out sample on almost all synthetic UMRs are close to 1, indicating global UMR expansion in knock-out sample. Strikingly, 16% of synthetic UMRs emerge in knockout sample, thus have no contribution from wild-type sample at all. In contrast, only 4% of synthetic UMRs disappear completely in the knock-out sample. Furthermore, each of the 508 synthetic UMRs has multiple wild-type UMRs separated by methylated CpGs, which are eroded in the knock-out sample such that these multiple UMRs are connected to a longer UMR. In contrast, only 177 wild-type UMRs are broken down into multiple UMRs in knockout sample.

To test if a given synthetic UMR is differentially methylated, we compared the mean methylation ratio of the synthetic UMR between wild-type and knock-out samples using MOABS with permutation FDR at 0.2%. The results indicate that 14% of synthetic LUMRs are differentially methylated between wild-type and knock-out samples.

Analysis of Oncomine-AML genes

We used Oncomine (Compendia Biosciences Ann Arbor, MI USA) to assess the enrichment of Canyon-associated genes expressed in WT murine HSCs (FPKM > 1) in patient signatures of genes over-expressed in Leukemic disease vs. normal bone marrow. Oncomine assesses overlap significance with Fisher's exact test. Our threshold criteria were Odds Ratio ≥ 1.8 and p-value < 1E-5. To address the challenge of cross-species comparison as well as the inherent technical limitations of comparing next generation sequencing data to that derived from legacy microarray technologies, we limited our analysis to signatures derived from the two most recent 3'IVT Affymetrix expression arrays represented in Oncomine, hgu133a and hgu133plus2, which interrogate 12,624 and 19,574 unique genes, respectively.

Generation of random gene sets (each approximating the number of expressed Canyon genes) and mapping of mouse to human gene homologs were performed in R with the Bioconductor package 'annotationTools' (Kuhn 2008). Simulated Canyons Genes represent randomly sampled genes with promoter enrichment or depletion of H3K27me3 and/or H3K4me3 histone modifications proportionate to the distributions observed in WT HSC Canyons as determined by ChIP-seq (Fig 2A). The gene set distribution was as follows: K4+K27+ (45.92%), K4-K27+ (6.97%), K4+K27− (46.56%), and K4-K27− (0.54%). Bivalent promoters (K4+K27+) required an overlap of at least 1 nucleosome length (~ 146 bp) between H3K4me3 and H3K27me3 peaks. Random unmethylated promoters were sampled from promoters (excluding Canyon genes) with mean CpG methylation level < 10% in WT HSC. Promoters regions were defined as ±1 kb relative to TSS in Refseq transcripts. Random expressed genes were sampled from genes with FPKM > 1 in WT HSC. All gene sets and Oncomine signatures represented in the analysis are provided (Supplementary Table 8).

Analysis of TCGA-AML genes

We downloaded the RNA-Seq data of AML patients from The Cancer Genome Atlas (TCGA) Data Portal (https://tcga-data.nci.nih.gov/tcga/) and performed the preprocessing; log2 transformation, orthologous gene mapping, and data filtering out genes over 20% missing values. In the process human gene symbols were mapped to those of mouse using the human and mouse homology information in Mouse Genome Informatics (http://informatics.jax.org). Finally, we selected the gene expression data with 14701 genes on 167 patients. Two sample t-test was applied to identify the significantly differentially expressed genes between two groups based on Dnmt3a mutation status. 1760 signature genes were selected (p-val < 0.05). BRB-ArrayTools and R language (http://www.r-project.org) were primarily used for statistical analysis of gene expression data[45]. Cluster analysis was performed using the software programs Cluster and Heatmap was generated by Treeview[46]. We assessed the enrichment of expressed Canyon genes (FPKM >1) in Dnmt3a mutation signatures using Hypergeometric test in R language (http://www.r-project.org).

Analysis of Canyon genes in Cancer Cell Line Encyclopedia (CCDE)

Gene expression in 947 cancer cell lines from cell line encyclopedia (CCDE: GSE36139) were used for hierarchical clustering of Canyon genes. Cluster analysis was performed using the software programs Cluster and Heatmap was generated by Treeview[46].

Data

All the data sets can be downloaded with GEO accession number GSE49191 (www.ncbi.nlm.nih.gov/geo/).

Track hub http://dldcc-web.brc.bcm.edu/lilab/benji/canyon.tracks.txt

This file contains HSC WGBS, RNA-Seq, ChIP-Seq, CMS-Seq and Canyon browser track. To upload the Data S1, go to the UCSC genome browser page for the mouse genome mm9, select Track hub under myData tab, then copy and paste this URL (http://dldcc-web.brc.bcm.edu/lilab/benji/canyon.tracks.txt) and hit add Hub button.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank Taiping Chen, J.P. Issa, John Gilbert, and members of the Goodell Lab for helpful discussions. This work was supported by NIH grants, AG036562, CA126752, DK092883, CA125123, DK084259, AI07495, the Ellison Medical Foundation, CPRIT grant RP110028 and the Samuel Waxman foundation; to W.L.: RP110471, NIH

R01HG007538 and DOD PC094421, and to A.R. CA151535, CIRM RM1-01729, and Leukemia and Lymphoma Society award 6187-12. Y.H. and M.K. are supported by Leukemia and Lymphoma Society fellowships.

References

1. Bird A, Taggart M, Frommer M, Miller OJ, Macleod D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell. 1985; 40:91–99. [PubMed: 2981636]
2. Irizarry RA, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2009; 41:178–186. [PubMed: 19151715]
3. Ley TJ, et al. DNMT3A mutations in acute myeloid leukemia. N Engl J Med. 2010; 363:2424–2433. [PubMed: 21067377]
4. Yan XJ, et al. Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. Nat Genet. 2011; 43:309–315. [PubMed: 21399634]
5. Delhommeau F, et al. Mutation in TET2 in myeloid cancers. N Engl J Med. 2009; 360:2289–2301. [PubMed: 19474426]
6. Abdel-Wahab O, et al. Genetic characterization of TET1, TET2, and TET3 alterations in myeloid malignancies. Blood. 2009; 114:144–147. [PubMed: 19420352]
7. Lister R, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009; 462:315–322. [PubMed: 19829295]
8. Stadler MB, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011; 480:490–495. [PubMed: 22170606]
9. Ziller MJ, et al. Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. PLoS Genet. 2011; 7:e1002389. [PubMed: 22174693]
10. Hodges E, et al. Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. Mol Cell. 2011; 44:17–28. [PubMed: 21924933]
11. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. J Mol Biol. 1987; 196:261–282. [PubMed: 3656447]
12. Lowe CB, Bejerano G, Haussler D. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc Natl Acad Sci U S A. 2007; 104:8005–8010. [PubMed: 17463089]
13. Hendrix DA, Hong JW, Zeitlinger J, Rokhsar DS, Levine MS. Promoter elements associated with RNA Pol II stalling in the Drosophila embryo. Proc Natl Acad Sci U S A. 2008; 105:7762–7767. [PubMed: 18505835]
14. Bellen HJ, et al. The Drosophila gene disruption project: progress using transposons with distinctive site specificities. Genetics. 2011; 188:731–743. [PubMed: 21515576]
15. Akalin A, et al. Transcriptional features of genomic regulatory blocks. Genome Biol. 2009; 10:R38. [PubMed: 19374772]
16. Kikuta H, et al. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. Genome Res. 2007; 17:545–555. [PubMed: 17387144]
17. Whyte WA, et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. 2013; 153:307–319. [PubMed: 23582322]
18. Wilson NK, et al. Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. Cell Stem Cell. 2010; 7:532–544. [PubMed: 20887958]
19. Bernstein BE, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006; 125:315–326. [PubMed: 16630819]
20. Stadler MB, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011; 480:490–495. [PubMed: 22170606]
21. Coolen MW, et al. Consolidation of the cancer genome into domains of repressive chromatin by long-range epigenetic silencing (LRES) reduces transcriptional plasticity. Nat Cell Biol. 2010; 12:235–246. [PubMed: 20173741]

22. Bert SA, et al. Regional activation of the cancer genome by long-range epigenetic remodeling. Cancer Cell. 2013; 23:9–22. [PubMed: 23245995]

23. Patel JP, et al. Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. N Engl J Med. 2012; 366:1079–1089. [PubMed: 22417203]

24. Pastor WA, Aravind L, Rao A. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. Nat Rev Mol Cell Biol. 2013; 14:341–356. [PubMed: 23698584]

25. Inoue A, Zhang Y. Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. Science. 2011; 334:194. [PubMed: 21940858]

26. Pastor WA, et al. Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. Nature. 2011; 473:394–397. [PubMed: 21552279]

27. Booth MJ, et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. Science. 2012; 336:934–937. [PubMed: 22539555]

28. Challen GA, et al. Dnmt3a is essential for hematopoietic stem cell differentiation. Nat Genet. 2012; 44:23–31. [PubMed: 22138693]

29. Roller A, et al. Landmark analysis of DNMT3A mutations in hematological malignancies. Leukemia. 2013

30. Jones PA, Baylin SB. The epigenomics of cancer. Cell. 2007; 128:683–692. [PubMed: 17320506]

31. Ganguly B, Foi A, Doctorovich F, B KD, T GR. 5-Hy-droxy-3-methyl-5-phenyl-4,5-di-hydro-1H-pyrazole-1-carbothio-amide. Acta Crystallogr Sect E Struct Rep Online. 2011; 67:o2777.

32. Xie W, et al. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell. 2013; 153:1134–1148. [PubMed: 23664764]

33. Ko M, et al. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. Nature. 2010; 468:839–843. [PubMed: 21057493]

References for Methods

34. Goodell MA, Brose K, Paradis G, Conner AS, Mulligan RC. Isolation and functional properties of murine hematopoietic stem cells that are replicating in vivo. J Exp Med. 1996; 183:1797–1806. [PubMed: 8666936]

35. Mayle A, Luo M, Jeong M, Goodell MA. Flow cytometry analysis of murine hematopoietic stem cells. Cytometry A. 2013; 83:27–37. [PubMed: 22736515]

36. Gu H, et al. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc. 2011; 6:468–481. [PubMed: 21412275]

37. Huang Y, Pastor WA, Zepeda-Martinez JA, Rao A. The anti-CMS technique for genome-wide mapping of 5-hydroxymethylcytosine. Nat Protoc. 2012; 7:1897–1908. [PubMed: 23018193]

38. Dahl JA, Collas P. A rapid micro chromatin immunoprecipitation assay (microChIP). Nat Protoc. 2008; 3:1032–1045. [PubMed: 18536650]

39. Xi Y, Li W. BSMAP: whole genome Bisulfite Sequence MAPping program. BMC Bioinformatics. 2009; 10:232. [PubMed: 19635165]

40. Lin X, et al. BSeQC: quality control of bisulfite sequencing experiments. Bioinformatics. 2013

41. Zhang Y, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008; 9:R137. [PubMed: 18798982]

42. Li R, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009; 25:1966–1967. [PubMed: 19497933]

43. He, Y.-F. et, al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science (New York, NY). 2011; 333:1303–1307.

44. Robinson M, McCarthy D, Smyth G. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009

45. Simon R, et al. Analysis of gene expression data using BRB-ArrayTools. Cancer Inform. 2007; 3:11–17. [PubMed: 19455231]

46. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998; 95:14863–14868. [PubMed: 9843981]

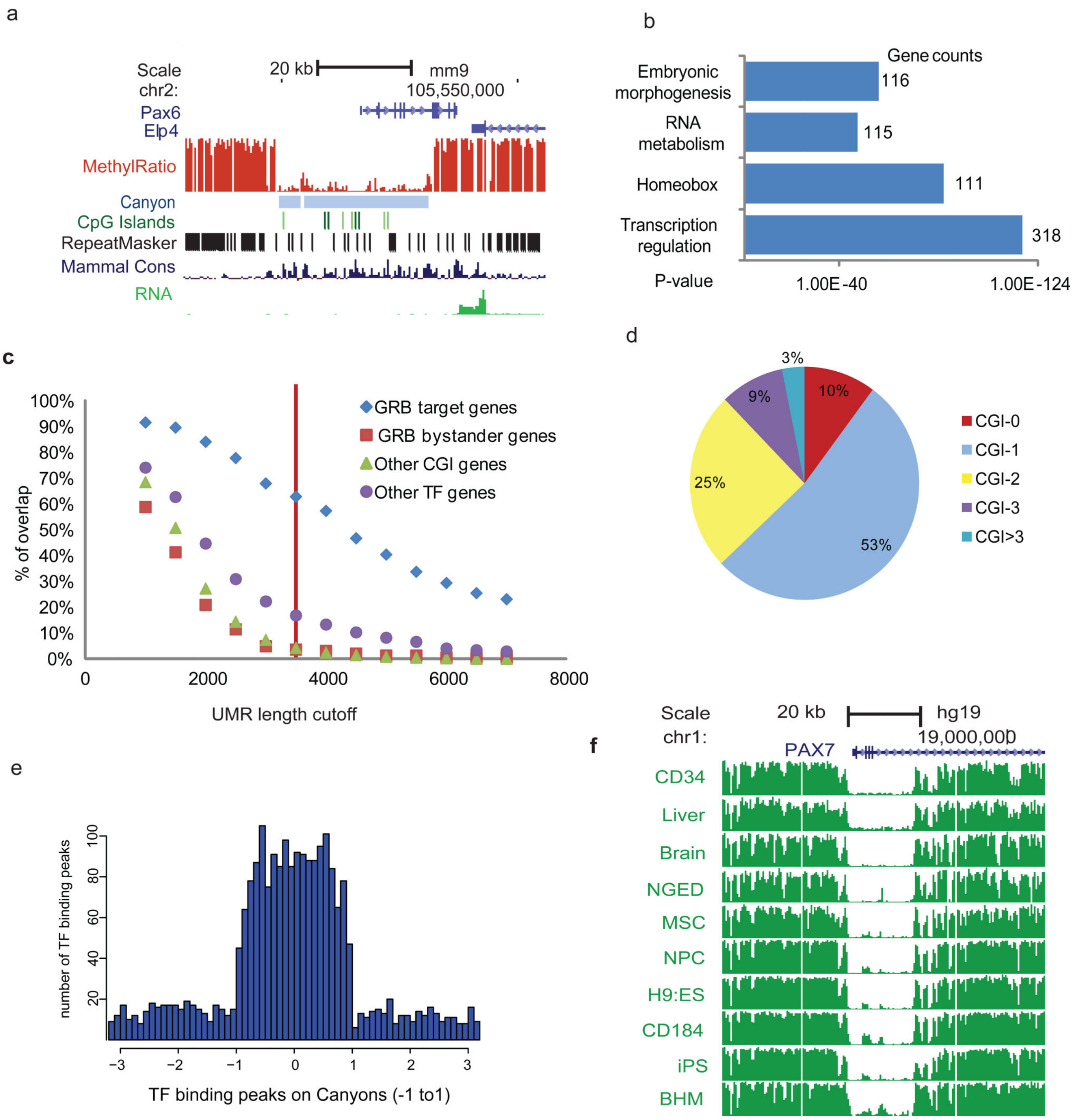

Figure 1. Large undermethylated Canyons revealed by WGBS

(**a**) UCSC genome browser track depicts methylation profile across the *Pax6* gene in murine HSCs. Methylation ratios from 0% to 100%, for individual CpG sites are shown in red. The identified Undermethylated regions (UMRs) (≤10% methylation) are indicated by blue bars, while the CpG islands are indicated in green, repeats are marked in black, and mammalian conservation is shown in dark blue. RNA-seq expression is shown at bottom in green (the *Pax6* promoter is in the center of the Canyon and has no RNAseq signal; the signal on the right of the plot comes from the 3' end of the adjacent gene which is transcribed toward

Pax6). (**b**) Gene ontology analysis of Canyon-associated genes. Ontology terms are shown on the y-axis; p-value for each category based on functional studies is graphed along the x-axis. (**c**) Overlap of four gene groups [15] using different UMR-length cutoffs. GRB targets genes are predicted regulatory targets of the highly conserved non-coding elements in genomic regulatory blocks (GRBs). Bystander genes are contained within GRBs but under distinct control. Other CGI genes overlap with CGI, but are not associated with GRBs, and the Other TF genes are transcription factors not associated with GRBs. The x-axis indicates the length cutoff of UMRs and y-axis indicates the percent of UMR-overlapping genes relative to all genes in each respective group. (**d**) The proportion of Canyons that contain the indicated numbers of CGIs. (**e**) Position of binding peaks for 10 TFs *(SCL/TAL1*, *LYL1*, *LMO2*, *GATA2*, *RUNX1*, *MEIS1*, *PU.1*, *ERG*, *FLI-1*, and *GFI1B*) across Canyons. The normalized Canyons are indicated with a position of 0 representing the Canyon centers, and positions ±1 representing the Canyon edges, as indicated by the blue bar. (**f**) *Pax7*-asscociated Canyons in human cells; data from the human Epigenome Atlas project. CD34: Mobilized CD34+ primary cells, Liver: Adult Liver, Brain: Brain Germinal Matrix, NGED: Neurosphere cultured cells– Ganglionic Eminence Derived, MSC: Human ES cell (H1)–derived Mesenchymal stem cells, NPC: H1-derived Neuronal Progenitor cultured cells, Human ES cells (H9), CD184: CD184+ Endoderm cultured cells, iPS: iPS DF 6.9 cell line, BHM: Brain Hippocampus Middle. Data obtained from NIH Roadmap Epigenomics Mapping Consortium (www.roadmapepigenomics.org).

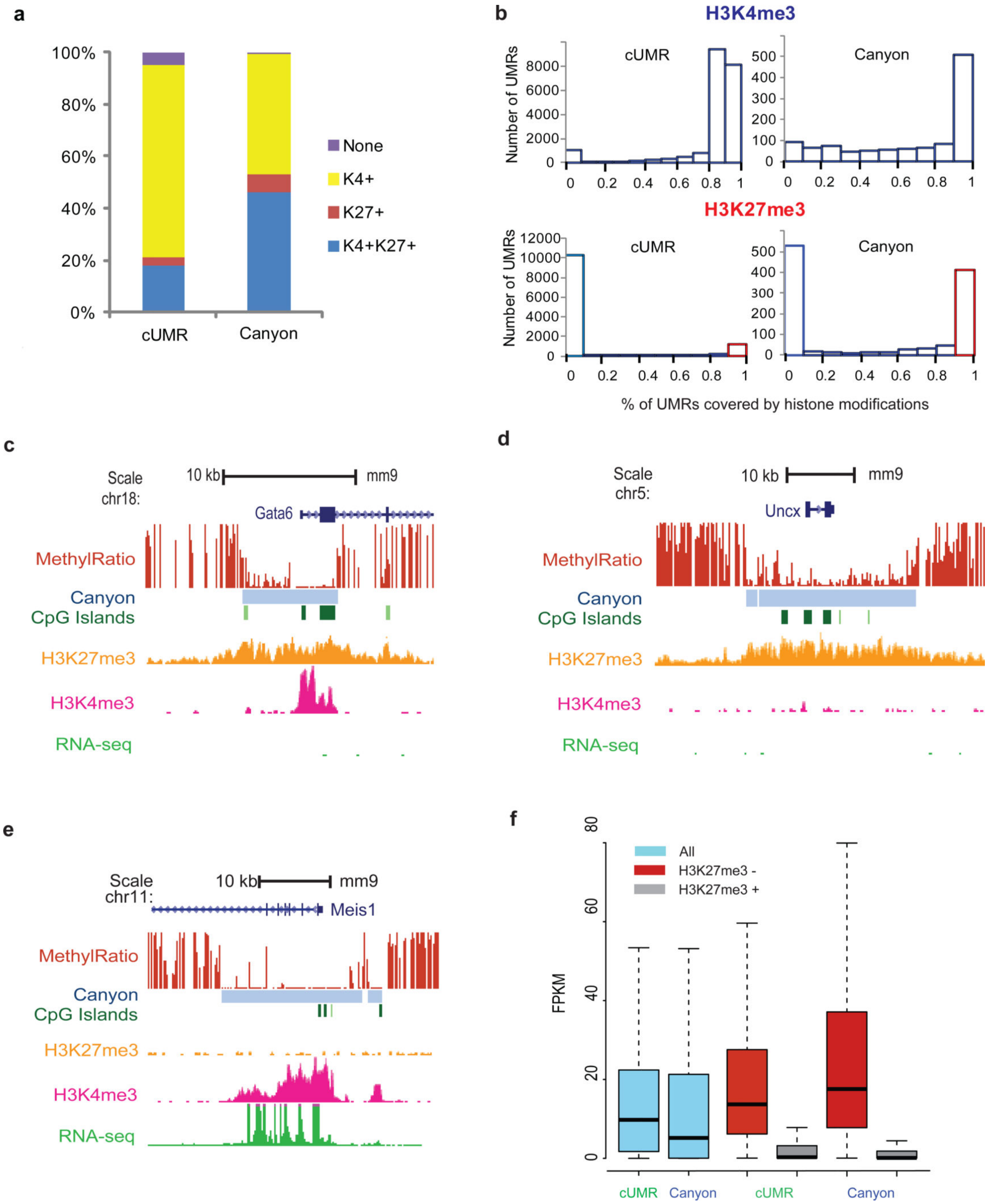

Figure 2. Histone modification and gene expression of Canyon-associated genes

(**a**) Proportion of UMRs or Canyons largely coated with the indicated histone marks. (**b**) Bar graph showed the UMR coverage by histone modifications. The number of UMRs with a given percent coverage of specific histone mark was plotted. (**c**) UCSC genome browser track depicting DNA methylation (red), H3K27me3 (yellow) and H3K4m3 (pink) and RNAseq data (light green) across *Gata6* gene (**d**) Depiction of the *Meis1* locus and (**e**) *Uncx* locus with the tracks as in (c). (**f**) Box plots show the distribution of average expression levels of cUMR- and Canyon-associated genes. The bottom and top of the box represent 25th and 75th percentile, while whiskers represent extension of 1.5 inter quartile range from the box. The horizontal line indicates the median value.

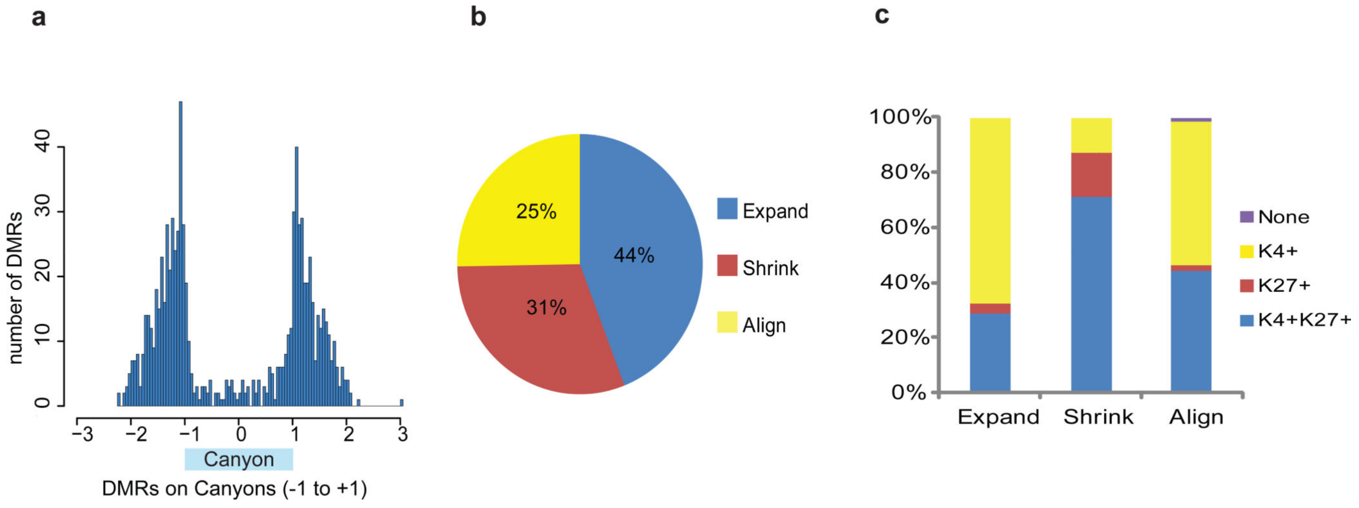

Figure 3. Erosion of Canyon borders in *Dnmt3a* KO HSC

(**a**) Position of differentially methylated regions (DMRs) comparing WT and *Dnmt3a* KO HSC on Canyons. The DMR position on Canyons is defined as relative distance between DMR center and Canyon center. The normalized Canyons are indicated with a position of 0 representing the Canyon centers, and positions ±1 representing the Canyon edges, as indicated by the blue bar. (**b**) Pie chart shows Canyon size dynamics in *Dnmt3a* KO HSC. (**c**) Distribution of histone marks associated with Canyon dynamics in *Dnmt3a* KO HSC (Canyons as defined in WT HSCs).

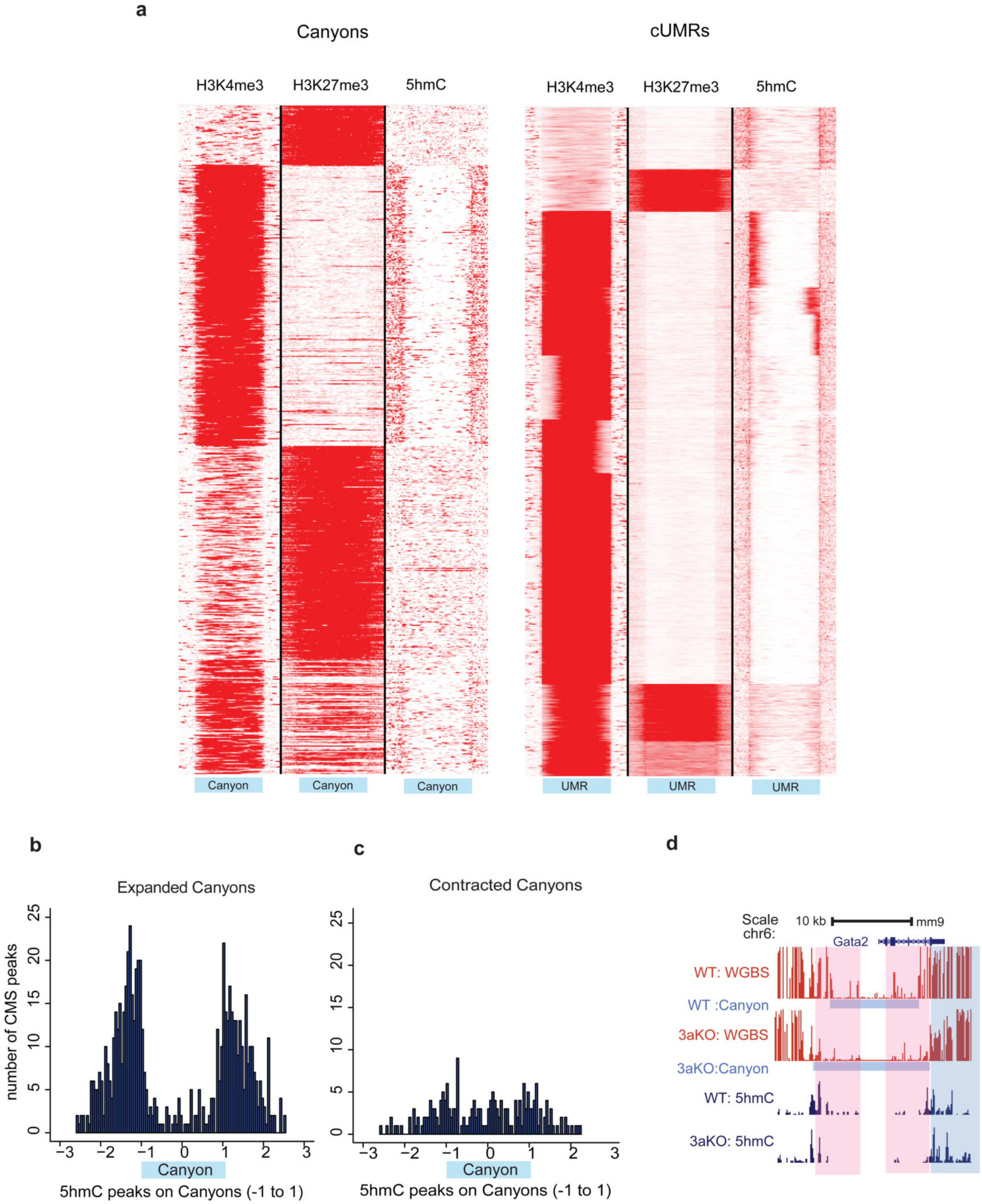

Figure 4. Histone and 5hmC distribution on Canyons and cUMRs

(**a**) Heatmap and profile of H3K4me3 and H3K27me3 around all the Canyons and cUMRs. All Canyons are normalized to same length. ±1 represents the boundary of Canyons and UMRs. Red represents high intensity and White represents no signal. Light blue bar shows normalized Canyon position. (**b**) Position of 5hmC peaks in WT HSCs on Canyons that expand in *Dnmt3a* KO HSCs. The normalized Canyons are indicated with a position of 0 representing the Canyon centers, and positions ±1 representing the Canyon edges, as indicated by the blue bar. (**c**) Position of 5hmC peaks in WT HSCs on contracting Canyons.

(**d**) UCSC genome browser track depicts methylation profiles and 5hmC peaks across the *Gata2* gene in WT and *Dnmt3a* KO HSCs. The Pink box indicates a methylation-depleted region with decreased 5hmC signal, and the blue box indicates a slightly methylation-decreased region with increased 5hmC signal.

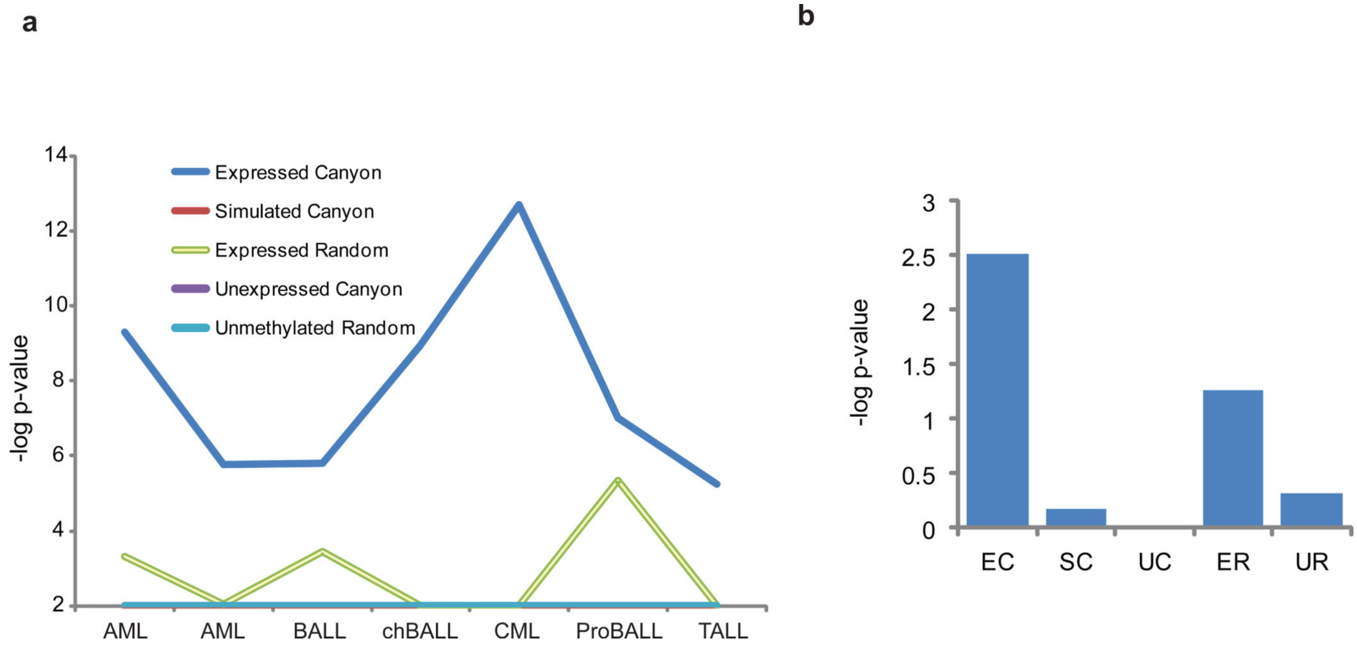

Figure 5. Aberrant expression of Canyon genes in hematologic malignancies

(**a**) Graph shows the association of Canyon genes expressed in WT murine HSCs (FPKM > 1) with Leukemia patient gene expression signatures in Oncomine (database version 4.4.3). Applying a stringent threshold cutoff (OR ≥ 1.8 p-value < 1.0E-55), we identified 7 signatures representing the top 10% of genes over-expressed in disease vs. normal bone marrow. Their enrichment was then compared to that of four controls randomly sampled from WT HSC: expressed genes, unexpressed genes, simulated Canyon genes, and genes outside of Canyons lacking promoter CpG methylation (Supplementary Table 8). Lines represent the negative log-transformed p-values for association of the indicated signature with expressed Canyon genes and controls. Note, Oncomine does not report associations with p-values < 0.01. AML, Acute Myeloid Leukemia; B-ALL, B Cell Acute Lymphoblastic Leukemia; ch, childhood; Pro-B ALL, Pro-B Cell Acute Lymphoblastic Leukemia; T-ALL, T-Cell Acute Lymphoblastic Leukemia. (**b**) Bar graph shows the association Canyon genes expressed in WT murine HSCs (FPKM > 1) with AML patients with Dnmt3a mutation differential gene expression signatures in TCGA data. Applying two samples t-test, we identified differentially expressed genes between AML patients with and without *DNMT3A* mutation (p<0.05). Their enrichment was compared with same control gene groups used for (Fig.4a). EC: Expressed Canyon, UC: Unexpressed Canyon, ER: Expressed Random, SC: Simulated Canyon, UR: Unmethylated Random. Y-axis represents the negative log transformed p-values.