Large conserved domains of low DNA methylation maintained by Dnmt3a

Mira Jeong[1,12], Deqiang Sun[2,12], Min Luo[1,12], Yun Huang[3], Grant A Challen[1,11], 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[7,8], Lucy A Godley[9], Gretchen J Darlington[10], Anjana Rao[3], Wei Li[2,13] & Margaret A Goodell[1,13]

Gains and losses in DNA methylation are prominent features of mammalian cell types. To gain insight into the mechanisms that promote shifts in DNA methylation and contribute to changes in cell fate, including malignant transformation, we performed genome-wide mapping of 5-methylcytosine and 5-hydroxymethylcytosine in purified mouse 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. About half of the genes in these methylation canyons are coated with repressive histone marks, whereas the remainder are covered by activating histone marks and are highly expressed in hematopoietic stem cells (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 new epigenetic landscape we describe may provide a mechanism for the regulation of hematopoiesis and may contribute to leukemia development.

In the mammalian genome, the majority of cytosines adjacent to guanines (constituting CpGs) are methylated (5mC), except in gene regulatory regions, where they are often clustered and are unmethylated (CpG islands, CGIs)[1]. Although regions with low levels of CpG methylation are considered generally permissive for gene expression when present in promoter regions, it is still only poorly understood how DNA methylation patterns vary among normal cell types, how they are added and erased, and how they influence gene expression. Whereas 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 hypomethylation in malignant cells[2]. However, most understanding of change in DNA methylation patterns has come from a limited analysis of cell lines, tissues of heterogeneous composition and 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* (refs. 3–6) has underscored the importance of DNA methylation in the 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, HSCs, and to investigate the factors that affect alterations in DNA methylation and gene expression.

RESULTS

DNA methylome in mouse HSCs

We performed whole-genome bisulfite sequencing (WGBS) with two biological replicates on purified mouse HSC side population cells that were also negative for lineage marker and positive for c-Kit, Sca-1 and CD150 (Online Methods). We generated a total of 1,121 million reads, of which 80.2% were successfully aligned to either strand of the reference genome (mm9), achieving a combined average coverage of 40× (**Supplementary Table 1**). For the two replicates, the data were highly reproducible, with a correlation coefficient of greater than 0.99 for methylation ratios across the genome. In general, the HSC methylome was similar to that of other mammalian cells[7,8]. Levels of DNA methylation were low at CGIs and promoters, and were higher in gene bodies and repetitive elements (**Supplementary Fig. 1**). In addition, non-CpG methylation

[1]Stem Cells and Regenerative Medicine Center, Department of Pediatrics and Molecular & Human Genetics, Baylor College of Medicine, Houston, Texas, USA. [2]Division of Biostatistics, Dan L. Duncan Cancer Center and Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, Texas, USA. [3]Division of Signaling and Gene Expression, La Jolla Institute for Allergy and Immunology, La Jolla, California, USA. [4]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA. [5]Department of Hematology, Cambridge Institute for Medical Research and Wellcome Trust and Medical Research Council Cambridge Stem Cell Institute, Cambridge University, Cambridge, UK. [6]Department of Systems Biology, Division of Cancer Medicine, University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [7]Department of Pathology, Baylor College of Medicine, Houston, Texas, USA. [8]Department of Biology & Biochemistry, University of Houston, Houston, Texas, USA. [9]Department of Medicine, University of Chicago, Chicago, Illinois, USA. [10]Huffington Center for Aging, Baylor College of Medicine, Houston, Texas, USA. [11]Present address: Department of Internal Medicine, Washington University at St. Louis, St. Louis, Missouri, USA. [12]These authors contributed equally to this work. [13]These authors jointly directed this work. Correspondence should be addressed to M.A.G. (goodell@bcm.edu) or W.L. (wl1@bcm.edu).

Received 26 July; accepted 1 November; published online 24 November 2013; doi:10.1038/ng.2836

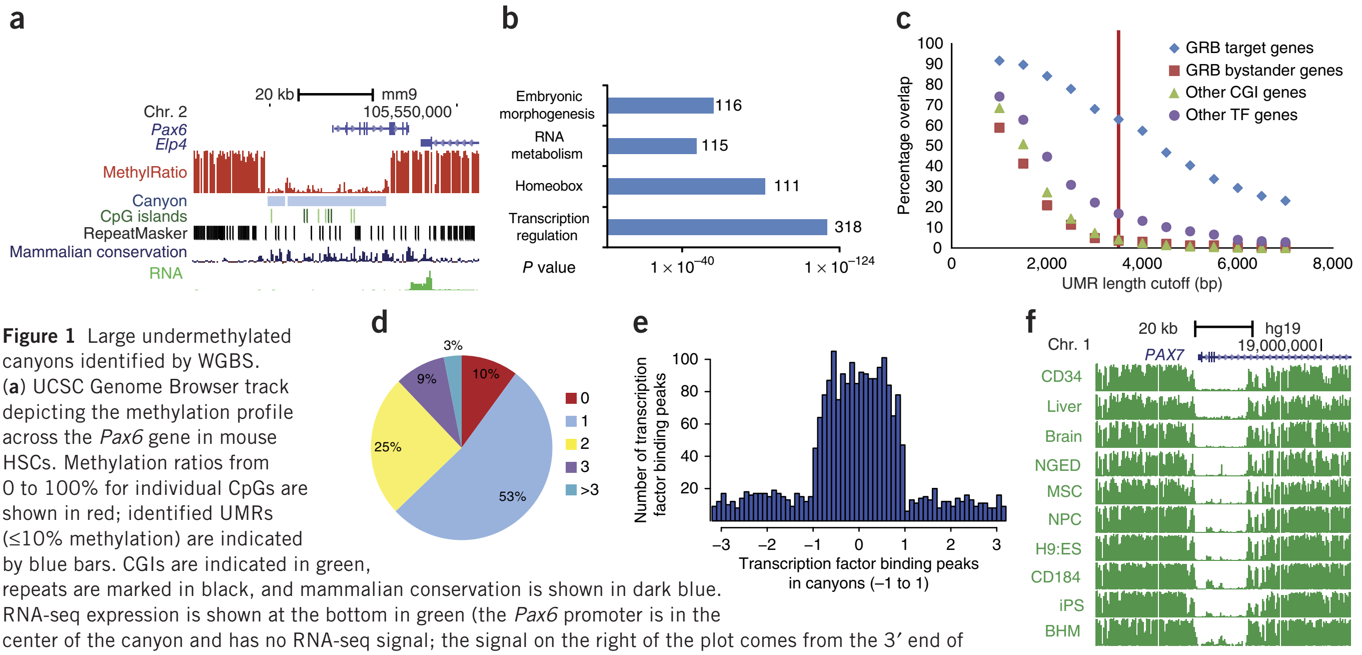


Figure 1 Large undermethylated canyons identified by WGBS. (**a**) UCSC Genome Browser track depicting the methylation profile across the *Pax6* gene in mouse HSCs. Methylation ratios from 0 to 100% for individual CpGs are shown in red; identified UMRs (≤10% methylation) are indicated by blue bars. CGIs are indicated in green, repeats are marked in black, and mammalian conservation is shown in dark blue. RNA-seq expression is shown at the bottom in green (the *Pax6* promoter is in the center of the canyon and has no RNA-seq 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, and the *P* value for each category based on functional studies is graphed along the *x* axis. The number of genes in each category is indicated to the right of the corresponding bar. (**c**) Overlap of four gene groups[15] using different UMR length cutoffs. The vertical red line indicates the 3.5-kb size cutoff used to define canyons. GRB target genes are predicted regulatory targets of highly conserved noncoding elements. Bystander genes are contained within GRBs but are under distinct control mechanisms. Other CGI genes overlap with CGIs but are not associated with GRBs, and other TF genes are transcription factors not associated with GRBs. The *x* axis indicates the length cutoff for UMRs, and the *y* axis indicates the percent of UMR-overlapping genes relative to all genes in each respective group. (**d**) Proportion of canyons that contain the indicated numbers of CGIs. (**e**) Position of binding peaks for ten transcription factors (Scl (Tal1), Lyl1, Lmo2, Gata2, Runx1, Meis1, Pu.1, Erg, Fli-1 and Gfl1b) across canyons. Canyons are normalized to position 0, representing the canyon center, and positions −1 and 1 represent the canyon edges. (**f**) *Pax7*-asscoiated canyons in human cells. 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 were obtained from the NIH Roadmap Epigenomics Mapping Consortium (see URLs).

was infrequent (less than 1% CpH methylation), consistent with data from other non–embryonic stem (ES) cell types[9].

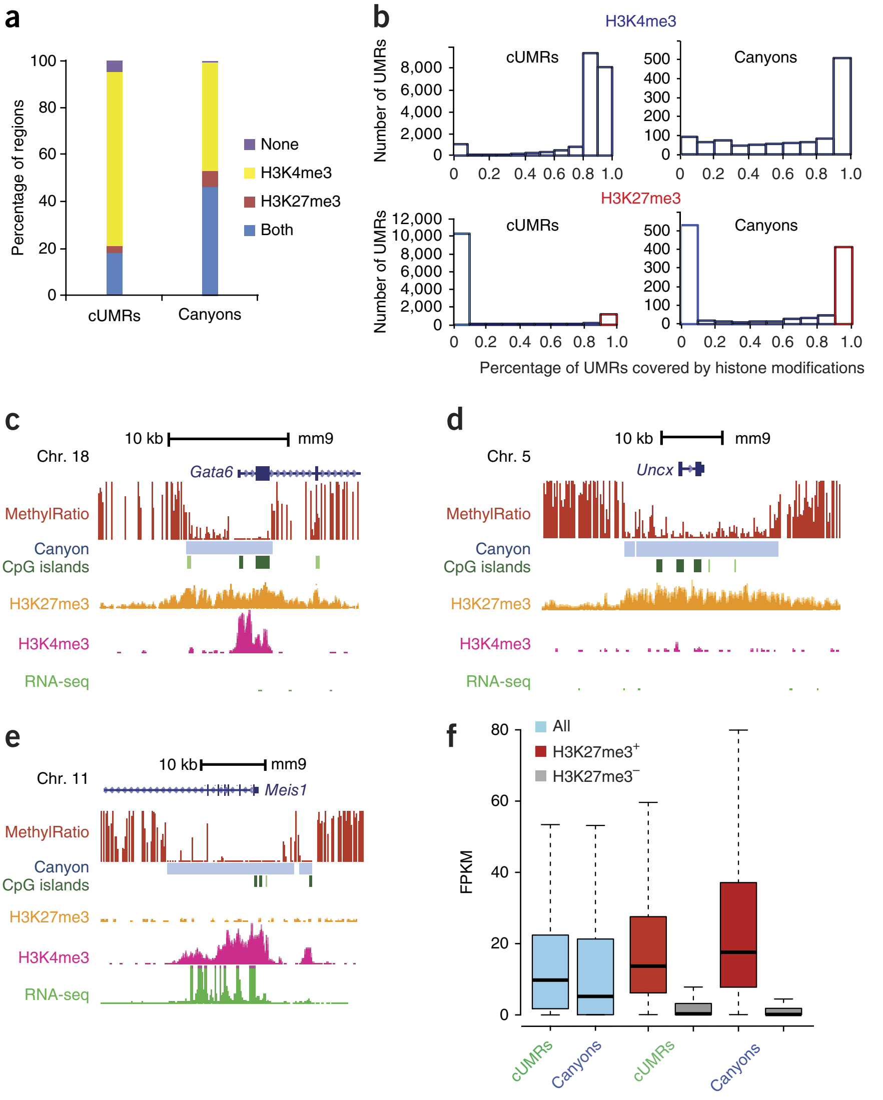
Canyons: large undermethylated genomic features

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 (HMM) to identify undermethylated regions (UMRs) with an average proportion of methylation of ≤10% (**Supplementary Table 2**), requiring at least five CpGs per kilobase to satisfy the permutation-based false discovery rate (FDR) of 5%. Using these criteria, we identified 32,325 UMRs in the mouse HSC methylome. Most UMRs were associated with promoters or gene bodies, and only 8.3% showed intergenic localization. In inspecting the UMR size distribution, we observed that a small portion of these regions were exceptionally large, with some extending over 25 kb, such as the UMR associated with the *Pax6* gene (**Fig. 1a**), which represents an expanse of unmethylated DNA that is considerably larger than that previously reported. In the genome landscape, these large methylation-depleted regions appeared as canyons cut into a plateau of high methylation, usually encompassing a single gene.

To determine whether these large UMRs represented a 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]; Online Methods). Using this criterion, we identified 1,104 methylation canyons representing 3.4% of all UMRs (**Supplementary Fig. 2** and **Supplementary Table 2**). To compare these regions with typical UMRs, we established a control group of 13,579 UMRs (cUMRs) that were longer than 1 kb but shorter than 3.5 kb. This control group excluded the smallest UMRs that tended 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, $P = 6.2 \times 10^{-123}$), as well as genes encoding a homeobox domain (111 genes, $P = 3.9 \times 10^{-85}$) (**Fig. 1b**, **Supplementary Fig. 3a** and **Supplementary Table 3**), comprising one of the most ancient gene families involved in the embryonic development of bilaterians. In contrast, the genes associated with cUMRs all gave nonsignificant *P* values (1.0) for these four gene ontology terms. Of the 20 largest canyons, 15 harbored homeobox-encoding genes (**Supplementary Fig. 3b**). These canyons typically extended well beyond the immediate regions containing these genes (**Supplementary Fig. 3c**–**e**). As a group, canyons were particularly highly conserved (**Supplementary Fig. 4a**) and were depleted of transposable elements and repeats (**Fig. 1a** and **Supplementary Fig. 4b**,**c**).

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

Figure 2 Histone modification and expression of canyon-associated genes. (**a**) Proportion of UMRs or canyons largely coated with the indicated histone marks. (**b**) Bar graphs showing UMR coverage by histone modifications. The number of UMRs with a given percent coverage with a specific histone mark was plotted. (**c**–**e**) UCSC Genome Browser tracks depicting DNA methylation ratios (red), H3K27me3 (yellow), H3K4me3 (magenta) and RNA-seq data (green) across *Gata6* (**c**), *Uncx* (**d**) and *Meis1* (**e**). (**f**) Box plots showing the distribution of average expression levels for cUMR- and canyon-associated genes. The bottom and top of each box represent the 25th and 75th percentiles, and whiskers represent extensions of 1.5 times the interquartile range from the box. The horizontal lines indicate the median value. FPKM, fragments per kb of exon per million reads.



We noted that the conservation and pattern of gene ontology enrichment for canyon-associated genes was similar to that described for a group of genes considered to be targets of highly conserved noncoding elements within large (~1-Mb) genomic regulatory blocks (GRBs)[15,16]. To systematically test whether these regions overlapped, we examined the relationship between UMR size and inclusion of GRB target genes. We plotted overlap in membership with GRB genes and three control gene groups against membership in UMR gene groups defined by different UMR length cutoffs (**Fig. 1c**). We found that the group of UMRs that were ≥3.5 kb in length overlapped with 67% of the GRB target genes, whereas the remaining 31,221 UMRs of ≤3.5 kb in length overlapped with only 27% of GRB target genes ($P = 2 \times 10^{-16}$). 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. Whereas CGIs were present in most canyons, 10% did not contain a classically defined CGI[11], and 53% contained a single CGI and were only covered by CGIs at a median of 26%. Therefore, the presence of a CGI cannot by itself explain these methylation lacunae (**Fig. 1d**). CGI shores, which comprise 2 kb of sequence on either side of a CGI, have been shown to exhibit the greatest variation in methylation across cell types[2]. Because most canyons contained one or more CGI, they would also harbor CGI-associated shores.

Recently, it was reported that there are large genomic domains called superenhancers that are occupied by master transcription factors and the mediator complex[17]. Although these domains have not been defined in HSCs by binding of mediator complex proteins, we reasoned that sites at which multiple transcription factors bind across several hemato poietic cell types would approximate such regions. To examine the relationship of such sites with canyons, we compared canyons with transcription factor binding sites identified from more than 150 chromatin immunoprecipitation and sequencing (ChIP-seq) data sets across a variety of blood lineages (>10)[18]. Interestingly, we found significant enrichment ($P = 2 \times 10^{-16}$) for peaks representing binding of ten HSC pluripotency-associated transcription factors, not only in small cUMR regions, but also across the entirety of canyons, in comparison with surrounding regions (**Fig. 1e** and **Supplementary Fig. 5a**–**c**).

Methylation canyons are conserved among cell types and species

To determine whether canyons are stable features or vary by cell type, we identified canyons in ES cell methylome data[8] using the same criteria as for HSCs. Of 839 ES cell canyons (**Supplementary Table 4**), 82% (688) were largely shared by both cell types, although there was variation in the position of edges, region length and average methylation levels (**Supplementary Fig. 6a**–**e**). Similarly, many canyons identified in mouse HSCs could be identified in human hematopoietic progenitors and differentiated progeny (**Supplementary Fig. 6f**)[10] and non-hematopoietic cells, with minimal variation between cell types (**Fig. 1f** and **Supplementary Fig. 6g**). We found that 72–80% of canyons defined in mouse ES cells overlapped with canyons in methylome data for a variety of human cell types (**Supplementary Fig. 6h**). These findings establish that methylation canyons are a distinct genomic feature that is stable, albeit with subtle differences, across cell types and species. Whereas most canyons contain CGIs and shores, methylation within canyons varies within a limited range, in contrast to the majority of shores found associated with CGIs excluded from canyons.

Expression of canyon genes is regulated by histone modifications

Low DNA methylation levels are usually associated with active gene expression. However, many canyon-associated genes are developmental regulators that are not known to have roles across many cell and tissue types; thus, we examined their regulatory features in more detail in the hematopoietic system. RNA sequencing (RNA-seq) data indicated that, of the 20 largest canyons, only 2 harbored highly expressed genes: *Hoxa9* and *Meis1*, which encode transcription factors critical for hematopoiesis and are frequently deregulated in leukemia (**Supplementary Fig. 3a**). To examine whether histone modifications could account for the lack of expression from other canyon-associated genes, we investigated activating trimethylation at lysine 4 of histone H3 (H3K4me3) and repressive trimethylation at lysine 27 of histone H3 (H3K27me3) in these regions by ChIP-seq. Whereas most cUMRs were associated with high H3K4me3 and low H3K27me3 levels, canyons showed a distinct bimodal distribution, with around half exhibiting high levels of H3K4me3 marks and half exhibiting high levels of H3K27me3 marks (**Fig. 2a**,**b**).

Of the mouse HSC canyons, 6% were only modified with H3K27me3, and 45.9% exhibited both H3K27me3 and H3K4me3 marks, similar to the 'bivalent domains' found in ES cells (for example, *Gata6*; **Fig. 2c**)[19]. H3K27me3 marks often covered the entire length of the canyon, such that canyon edges were aligned with the edges of the H3K27me3 peak, as in the *Uncx* gene (**Fig. 2d**). Similarly, the remaining H3K27me3-negative canyons were heavily coated by H3K4me3 marks, such as the *Meis1* gene (**Fig. 2e**). H3K27me3 marks were the defining feature for expression, as the genes associated with canyons that only had H3K4me3 marks were highly expressed, whereas the ones associated with H3K27me3-modified canyons showed low or no expression, regardless of their association with H3K4me3 marks. The median expression of genes associated with canyons that only had H3K4me3 marks was higher than for comparable cUMR genes (**Fig. 2f**). Although in HSCs we could not determine whether individual cells harbor both activating and repressive histone marks on the same allele, these data are consistent with a special epigenetic status for a certain subset of developmentally important genes, in which they exhibit activation-associated DNA methylation lacunae 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 from which these loci will be expressed during differentiation. The putative presence of these loci in HSCs suggests instead that they reflect a privileged epigenetic status or perhaps indicate differentiation history rather than future potential.

Canyons partially overlap with other low-methylation regions

Most studies of DNA methylation have focused mainly on CGIs, defined as being more than 300 bp in length and having a CG content of over 50%, which are unmethylated and are generally associated with promoters. Recent genome-wide approaches have identified additional regions with important alterations in methylation in cancer and cell fate decisions, such as CGI shores[2], partially methylated domains (PMDs)[7], low-methylated regions (LMRs)[20] and long-range epigenetic activation (LREA) or 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. Although these methylation canyons share many features with smaller UMRs, they represent only 3.4% of all UMRs and are distinct in their very low levels of methylation, their enrichment for homeobox-encoding genes, their overlap with GRB target genes, their stability among cell types and the bimodal distribution of H3K27me3 marks representing a distinct mode of regulation of gene expression (**Supplementary Table 5**).

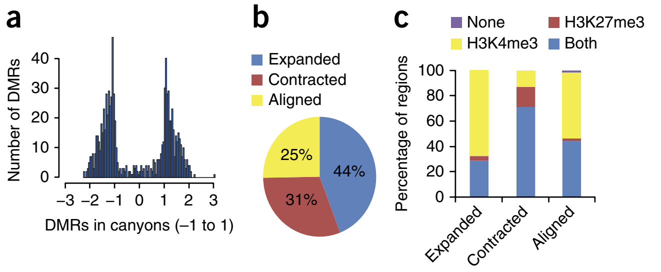


Figure 3 Erosion of canyon borders in *Dnmt3a*-null HSCs. (**a**) Positions of DMRs identified when comparing wild-type and *Dnmt3a*-null HSCs at canyons. The DMR position in a canyon is defined as the relative distance between the DMR center and the canyon center. Canyons are normalized to position 0, representing the canyon center, and positions −1 and 1 represent the canyon edges. (**b**) Pie chart showing canyon size dynamics in *Dnmt3a*-null HSCs. (**c**) Distribution of histone marks associated with canyon dynamics in *Dnmt3a*-null HSCs (canyons defined as in wild-type HSCs).

Methylation canyons are maintained by Dnmt3a

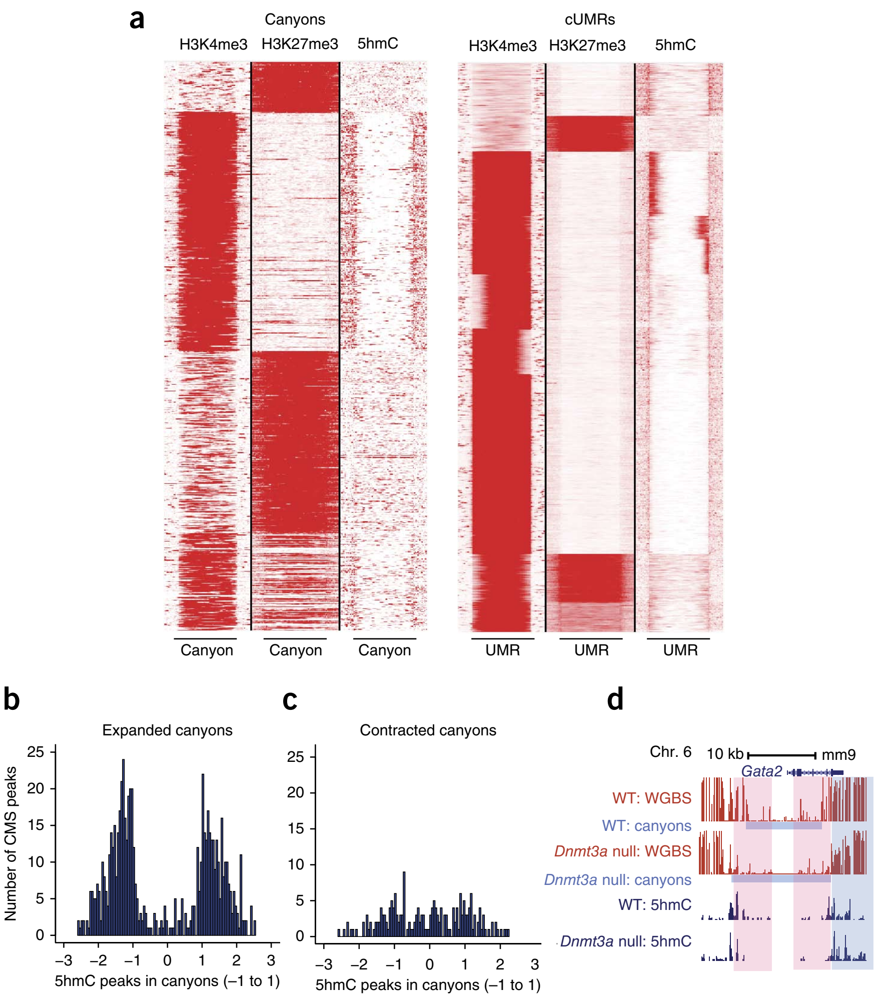
Because *DNMT3A* is mutated in a high frequency of human leukemias[23] and its loss in mouse HSCs leads to their expansion[24], we examined the impact of loss of Dnmt3a on canyon size. We compared all UMRs in HSCs with conditional inactivation of *Dnmt3a* to those in wild-type HSCs. With inactivation of *Dnmt3a*, we found that the edges of cUMRs and canyons became hotspots of differential methylation, whereas regions inside cUMRs and canyons were relatively resistant to variation (**Supplementary Fig. 7a**). Thirty percent of all differentially methylated regions (DMRs) in *Dnmt3a*-null cells were located at the edges of UMRs. This focused methylation loss at the edges of UMRs suggests that Dnmt3a normally acts to maintain regions of methylation at their boundaries (**Fig. 3a** and **Supplementary Fig. 7a**). In 44% of canyons, the edges were eroded, such that the regions increased in size, and 31% of canyons experienced hypermethylation at their edges, such that the regions decreased in size (25% experienced no significant change in size). Loss of methylation in *Dnmt3a*-null HSCs led to the addition of 861 new canyons, for a total of 1,787 canyons (**Fig. 3b** and **Supplementary Table 6**). Methylation in some regions that featured a cluster of canyons in wild-type 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 covered more than 50 kb, interrupted by short stretches with higher levels of methylation (**Supplementary Fig. 7b**).

The expansion and contraction of different canyons in the absence of *Dnmt3a* is reminiscent of the concomitant hyper- and hypomethylation that is observed in many malignant cells; thus, we considered whether other epigenetic mechanisms influence canyon behavior. We first examined the distribution of histone marks on expanding versus contracting canyons. Of the canyons defined in wild-type cells, those marked with only H3K4me3 were most likely to expand after *Dnmt3a* inactivation. In contrast, canyons marked only with H3K27me3 or with both signals were more likely to contract (**Fig. 3c**). These findings suggest that Dnmt3a is acting specifically to restrain canyon size where active histone marks (and active transcription) are already present.

Canyon borders are demarked by 5-hydroxymethylcytosine

We next considered whether the erosion of canyon edges was attributable to an active process. The Tet protein family may promote demethylation, as hydroxymethylated cytosine (5hmC) is not recognized by Dnmt1, leading to its replacement with unmethylated cytosine

Figure 4 Histone and 5hmC distribution on canyons and cUMRs. (**a**) Heat map and profile of H3K4me3 and H3K27me3 marks around all canyons and cUMRs. All canyons are normalized to the same length. Red represents high-intensity signal, and white represents no signal. (**b**,**c**) Positions of 5hmC peaks in wild-type HSCs in canyons that expanded (**b**) or contracted (**c**) in *Dnmt3a*-null HSCs. Canyons are normalized to position 0, representing the canyon center, and positions –1 and 1 represent the canyon edges. (**d**) UCSC Genome Browser track depicting methylation profiles and 5hmC peaks across the *Gata2* gene in wild-type (WT) and *Dnmt3a*-null HSCs. The pink box indicates a methylation-depleted region with decreased 5hmC signal, and the blue box indicates a region with slightly decreased methylation and increased 5hmC signal.



during DNA replication[25,26]. WGBS cannot distinguish between 5mC and 5hmC, so we determined the genome-wide distribution of 5hmC in wild-type and *Dnmt3a*-null HSCs by using the cytosine-5-methylenesulphonate sequencing (CMS-seq) method[27]. In this method, sodium bisulfite treatment converts 5hmC to CMS, and CMS-containing DNA fragments are then immunoprecipitated using a CMS-specific antiserum and sequenced (**Supplementary Table 7**). Several sites with CMS signal were validated using oxidative bisulfite sequencing (oxBS-seq)[28], a technique based on quantitative sequencing of 5mC and 5hmC at single-base resolution (**Supplementary Fig. 8a**,**b** and **Supplementary Table 8**). Strikingly, 5hmC peaks were enriched specifically at the borders of both cUMRs and canyons (**Supplementary Fig. 8c**). In particular, expanding canyons, typically associated with the highest levels of H3K4me3 marks, were highly enriched at their edges for 5hmC signal (**Fig. 4a**,**b**). In contrast, contracting canyons, more likely to be associated with H3K27me3 marks, were depleted of 5hmC at their edges (**Fig. 4a**,**c**). An example of an expanding canyon involved the HSC master regulator *Gata2*, which showed 5hmC peaks at the canyon boundaries in wild-type HSCs and erosion of methylation edges in *Dnmt3a*-null cells (**Fig. 4d**). In regions where the methylation signal was completely depleted in *Dnmt3a*-null cells, 5hmC peaks disappeared altogether, consistent with loss of the 5mC substrate for hydroxylation. In regions where methylation levels were merely reduced, the 5hmC signal tended to increase (**Supplementary Fig. 9a**–**d**), suggesting unimpeded access to the DNA by the Tet proteins. We would expect that additional divisions of the *Dnmt3a*-null HSCs would result in the elimination of methylation at these sites, possibly contributing to further decline in the differentiation potential of these cells[24]. It is worth noting that the *DNMT3A* somatic mutations found in humans with acute myeloid leukemia (AML) are distinct from the *Dnmt3a*-null allele used here, with many cases being heterozygous for an allele with a specific point mutation encoding Arg882 in the catalytic domain[3] and others being compound heterozygous for likely inactivating mutations[29], such that the impact of *DNMT3A* mutations on canyon edges in different hematologic malignancies may be different than that seen here in *Dnmt3a*-null mouse HSCs. All three Tet family proteins are expressed in mouse HSCs; thus, we cannot determine which contributes to establishing and/or maintaining 5hmC signal. Going forward, the direct actions of Dnmt3a and specific Tet proteins on canyon edges ultimately need to be characterized through biochemical methods.

Canyon gene expression is associated with cancer

Aberrant hypermethylation in transformed cells has been thought to contribute to the development of malignancy[30], and both hyper- and hypomethylation are associated with transformed cells. Thus, we tested whether canyon-associated genes were likely to be associated with the development of hematologic malignancy. We used Oncomine to assess whether the canyon-associated genes expressed in wild-type HSCs were associated with the aberrant expression signatures identified in human leukemias. These canyon-associated genes were highly enriched in seven signatures of genes overexpressed in leukemias compared to normal bone marrow; in contrast, four sets of control genes were not enriched (**Fig. 5a** and **Supplementary Table 9**). Further, we used data from The Cancer Genome Atlas (TCGA) to test whether changes in the expression of canyon-associated genes were associated with *DNMT3A* mutation in AML in humans. Remarkably, we found that expressed canyon-associated genes were significantly enriched for differentially expressed genes in humans with AML with and without *DNMT3A* mutation ($P < 0.05$) (**Fig. 5b** and **Supplementary Table 9**). Overall, 76 expressed canyon-associated genes, including multiple *HOX* genes, were significantly changed in individuals with *DNMT3A* mutation ($P = 0.0031$) (**Supplementary Table 9**). Notably,

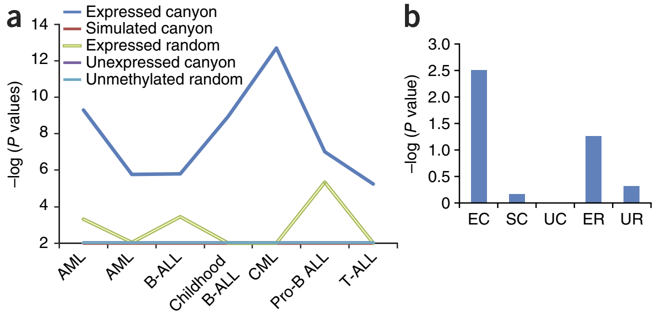


Figure 5 Aberrant expression of canyon-associated genes in hematologic malignancies. (**a**) Graph showing the association of canyon genes expressed in wild-type mouse HSCs (FPKM > 1) with gene expression signatures for humans with leukemia from Oncomine (database version 4.4.3). Applying a stringent criteria (odds ratio ≥ 1.8, $P < 1.0 \times 10^{-55}$), we identified seven signatures representing the top 10% of genes overexpressed in leukemias versus normal bone marrow. Their enrichment was then compared to that of four controls randomly sampled from wild-type HSCs: 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 that Oncomine does not report associations with $P < 0.01$. AML, acute myeloid leukemia; B-ALL, B cell acute lymphoblastic leukemia; pro-B ALL, pro-B cell acute lymphoblastic leukemia; T-ALL, T cell acute lymphoblastic leukemia. (**b**) Bar graph showing the association of canyon genes expressed in wild-type mouse HSCs (FPKM > 1) with differential gene expression signatures in humans with AML with *DNMT3A* mutation in TCGA data. Applying two-sample *t* tests, we identified differentially expressed genes in individuals with AML with and without *DNMT3A* mutation ($P < 0.05$). Their enrichment was compared with the same control gene groups used for **Figure 4a**. EC, expressed canyon; UC, unexpressed canyon; ER, expressed random; SC, simulated canyon; UR, unmethylated random. The *y* axis presents negative log-transformed *P* values.

whereas previous studies of differences in whole-transcriptome gene expression in leukemias with and without *DNMT3A* mutation did not identify any expression cluster associated with *DNMT3A* mutation[3], we identified two strong clusters from unsupervised clustering, with 80% of *DNMT3A*-mutant cases enriched in cluster A (**Supplementary Fig. 10**). The expressed canyon-associated genes identified here may be used as a unique gene expression signature to define the *DNMT3A* mutation status of patients. We further checked the expression of canyon-associated genes in various other cancer types by using data from a cancer cell line encyclopedia (Cancer Cell Line Encyclopedia (CCLE), a compilation of gene expression data from 947 human cancer cell lines). Canyon-associated genes expressed in HSCs were highly expressed or depleted in hematologic cancer cell lines, whereas unexpressed canyon-associated genes showed high expression in other cancer cell lines, which may reflect the original tissue specificity of canyon expression 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 levels of H3K4me3 trithorax-associated marks. 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 1,220 conserved genomic loci enriched for developmental regulators, which were also marked by either H3K4me3 or H3K27me3.

Active HSC canyons, containing genes involved in hematopoiesis and frequently dysregulated in leukemias, are particularly susceptible to loss of DNA methylation. This finding 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 with active chromatin marks. The insight that Tet proteins and Dnmt3a compete to maintain the status quo at the same loci in HSCs enables multiple scenarios to be envisioned in which the action of one protein or the other is reduced, either owing to attenuation of gene expression or mutation, leading to consequences for methylation, gene expression and developmental potential.

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

Mutations in *DNMT3A* and *TET2* have been linked to a similar spectrum of hematologic malignancies in humans[3–5,33]. Although the encoded proteins seem to oppose each other biochemically, genetically, mutations in these genes have a similar impact, impeding differentiation and promoting transformation. Although the precise mechanisms through which these activities occur are still unclear, the action of Dnmt3a and Tet proteins 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 changes in methylation and gene expression in both mouse models[28] and human samples[3] may reflect complex regulation at these loci, indicating the need to take multiple epigenetic factors into account as one seeks to understand the pathogenesis of these malignancies.

URLs. US National Institutes of Health Roadmap Epigenomics Mapping Consortium http://www.roadmapepigenomics.org/; MOABS (Model-Based Analysis of Bisulfite Sequencing), http://code.google.com/p/moabs/; TCGA Data Portal, https://tcga-data.nci.nih.gov/tcga/; Mouse Genome Informatics, http://informatics.jax.org/; R language, http://www.r-project.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. All data sets can be downloaded under Gene Expression Omnibus (GEO) accession GSE49191. UCSC Genome Browser tracks (mouse mm9) can be accessed from the hub at 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 tracks. To upload Data S1, go to the UCSC Genome Browser page for the mouse genome mm9, select "Track hub" under the "myData" tab and insert the URL http://dldcc-web.brc.bcm.edu/lilab/benji/canyon.tracks.txt, clicking on the "Add Hub" button.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank T. Chen, J.P. Issa, J. Gilbert and members of the Goodell laboratory for helpful discussions. This work was supported by US National Institutes of Health (NIH) grants AG036562, CA126752, DK092883, CA125123, DK084259 and AI07495, the Ellison Medical Foundation, Cancer Prevention Research Institute of Texas (CPRIT) grant RP110028 and the Samuel Waxman Foundation (M.A.G.): by CPRIT RP110471, NIH grant RO1HG007538 and DOD PC094421 (W.L.); and by grants CA151535 and CIRM RM1-01729 and Leukemia and Lymphoma Society award 6187-12 (A.R.). Y.H. and M.K. are supported by Leukemia and Lymphoma Society fellowships.

AUTHOR CONTRIBUTIONS

M.J., M.L., G.A.C., X.Z., Y.H., M.K., H.W., L.Y. and R.C. designed and performed experiments. M.J., D.S., M.L., G.A.C., B.R., L.C., S.-B.K., R.H., L.A.G., A.R., G.J.D., W.L. and M.A.G. analyzed data. M.J., D.S., M.L., G.A.C., B.R., J.-S.L., B.G., P.G., L.A.G., G.J.D., A.R., W.L. and M.A.G. wrote and edited the manuscript.

COMPETING FINALCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Bird, A., Taggart, M., Frommer, M., Miller, O.J. & Macleod, D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40**, 91–99 (1985).
2. Irizarry, R.A. *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41**, 178–186 (2009).
3. Ley, T.J. *et al. DNMT3A* mutations in acute myeloid leukemia. *N. Engl. J. Med.* **363**, 2424–2433 (2010).
4. Yan, X.J. *et al.* Exome sequencing identifies somatic mutations of DNA methyltransferase gene *DNMT3A* in acute monocytic leukemia. *Nat. Genet.* **43**, 309–315 (2011).
5. Delhommeau, F. *et al.* Mutation in *TET2* in myeloid cancers. *N. Engl. J. Med.* **360**, 2289–2301 (2009).
6. Abdel-Wahab, O. *et al.* Genetic characterization of *TET1*, *TET2*, and *TET3* alterations in myeloid malignancies. *Blood* **114**, 144–147 (2009).
7. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
8. Stadler, M.B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).
9. Ziller, M.J. *et al.* Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. *PLoS Genet.* **7**, e1002389 (2011).
10. Hodges, E. *et al.* Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol. Cell* **44**, 17–28 (2011).
11. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).
12. Lowe, C.B., Bejerano, G. & Haussler, D. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc. Natl. Acad. Sci. USA* **104**, 8005–8010 (2007).
13. Hendrix, D.A., Hong, J.W., Zeitlinger, J., Rokhsar, D.S. & Levine, M.S. Promoter elements associated with RNA Pol II stalling in the *Drosophila* embryo. *Proc. Natl. Acad. Sci. USA* **105**, 7762–7767 (2008).
14. Bellen, H.J. *et al.* The *Drosophila* gene disruption project: progress using transposons with distinctive site specificities. *Genetics* **188**, 731–743 (2011).
15. Akalin, A. *et al.* Transcriptional features of genomic regulatory blocks. *Genome Biol.* **10**, R38 (2009).
16. Kikuta, H. *et al.* Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* **17**, 545–555 (2007).
17. Whyte, W.A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
18. Wilson, N.K. *et al.* Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell* **7**, 532–544 (2010).
19. Bernstein, B.E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).
20. Stadler, M.B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011).
21. Coolen, M.W. *et al.* Consolidation of the cancer genome into domains of repressive chromatin by long-range epigenetic silencing (LRES) reduces transcriptional plasticity. *Nat. Cell Biol.* **12**, 235–246 (2010).
22. Bert, S.A. *et al.* Regional activation of the cancer genome by long-range epigenetic remodeling. *Cancer Cell* **23**, 9–22 (2013).
23. Patel, J.P. *et al.* Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. *N. Engl. J. Med.* **366**, 1079–1089 (2012).
24. Challen, G.A. *et al. Dnmt3a* is essential for hematopoietic stem cell differentiation. *Nat. Genet.* **44**, 23–31 (2012).
25. Pastor, W.A., Aravind, L. & Rao, A. TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nat. Rev. Mol. Cell Biol.* **14**, 341–356 (2013).
26. Inoue, A. & Zhang, Y. Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. *Science* **334**, 194 (2011).
27. Pastor, W.A. *et al.* Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473**, 394–397 (2011).
28. Booth, M.J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science* **336**, 934–937 (2012).
29. Roller, A. *et al.* Landmark analysis of *DNMT3A* mutations in hematological malignancies. *Leukemia* **27**, 1573–1578 (2013).
30. Jones, P.A. & Baylin, S.B. The epigenomics of cancer. *Cell* **128**, 683–692 (2007).
31. Long, H.K. *et al.* Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *eLife* **2**, e00348 (2013).
32. Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–1148 (2013).
33. Ko, M. *et al.* Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant *TET2*. *Nature* **468**, 839–843 (2010).

ONLINE METHODS

Hematopoietic stem cell purification and flow cytometry. For wild-type HSCs, cells were isolated from the whole bone marrow of femurs, tibias, pelvis bones and humeri from 12-month-old male C57BL/6 mice. Ten mice were used to purify HSCs, and biological replicates were performed with two separate pools of HCSs from different donors. *Dnmt3a*-null HSCs were purified from mice at the tertiary stage of serial transplantation because the phenotype resulting from loss of *Dnmt3a* is most strongly manifested at this stage[24]. Eighteen weeks after tertiary transplantation, donor cell–derived (CD45.2$^+$) HSCs were purified from four to eight transplanted mice per biological replicate. This timing allowed age-matched comparison to HSCs from 12-month-old wild-type mice.

HSCs from both wild-type and *Dnmt3a*-null mice were purified using the side population[34] strategy of Hoechst staining in combination with staining for surface markers[35]. Briefly, cells from whole bone marrow were resuspended in staining medium at 1 × 10^6 cells/ml and incubated with 5 mg/ml Hoechst 33342 (Sigma) for 90 min at 37 °C. For antibody staining, cells were suspended at a concentration of 1 × 10^8 cells/ml and incubated at 4 °C for 15 min with the desired antibodies. Magnetic enrichment was performed with biotin-conjugated antibody to c-Kit (eBioscience, 13-1171-82) and anti-biotin microbeads (Miltenyi Biotec) or with microbeads conjugated to antibody to mouse CD117 (Miltenyi Biotec) on an AutoMACS (Miltenyi Biotec). After enrichment, the positive cell fraction was labeled with antibodies to identify HSCs (SP$^+$Lineage$^-$Sca-1$^+$ (eBioscience, 25-5981-82) c-Kit$^+$ (eBioscience, 47-1171-82) CD150$^+$ (Biolegend, 115904). All antibodies were obtained from BD Biosciences or eBioscience and were used at 1:100 dilutions. Cell sorting was performed on a MoFlo cell sorter (Dako North America) or an Aria II (BD Biosciences), and analysis was performed on an LSRII (BD Biosciences). All mouse work was performed with approval from the Baylor College of Medicine Institutional Animal Care and Use Committee.

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

CMS technique for detection of 5-hydroxymethylcytosine. For CMS precipitation[37], 1.5 μg of fragmented genomic DNA was ligated with methylated adaptors and treated with sodium bisulfite (Qiagen). DNA was then denatured for 10 min at 95 °C (0.4 M NaOH, 10 mM EDTA), the reaction was neutralized by the addition of cold 2 M ammonium acetate, pH 7.0, and samples were incubated with antiserum to CMS 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. Antibody was precipitated with Protein G beads. Precipitated DNA was eluted with Proteinase K, purified by phenol-chloroform extraction and amplified by eight cycles of PCR using Pfu TurboCx Hot-Start DNA polymerase. DNA sequencing was carried out using Illumina/Solexa Genome Analyzer II and HiSeq sequencing systems.

Oxidative bisulfite sequencing. Genomic DNA was further purified by ethanol precipitation and micro Bio-Spin 6 column (Bio-Rad). Purified genomic DNA (250 ng) was denatured in 24 μl of 0.05 M NaOH at 37 °C for 30 min and 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 genomic DNA on ice for 1 h, with occasional vortexing. The mixture was purified by micro Bio-Spin 6 column. Non-oxidized and oxidized genomic DNA fractions were prepared using the MethylCode bisulfite conversion kit (Invitrogen). Locus-specific PCR was performed with the PyroMark PCR kit (Qiagen). Amplicons were pooled, and barcoded libraries were prepared with the TruSeq library preparation kit. Amplicon sequencing was performed on a MiSeq instrument (Illumina).

Computational analysis of oxidative bisulfite sequencing data. Bisulfite and oxidative bisulfite sequencing data were mapped against the mm9 reference genome using Bismark software[38] v0.6.4 (-q -n 2-chunkmbs 1028 bowtie-0.12.7). Subsequently, the number of reads containing converted cytosines 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 bisulfite sequencing and oxidative bisulfite sequencing, the percentage of hydroxymethylation was calculated by subtracting the observed methylation in oxidative bisulfite sequencing from the observed methylation in bisulfite sequencing.

RNA sequencing. For RNA-seq, ~70,000 HSCs were sorted into TRIzol from pools of cells. RNA was isolated with RNeasy Micro columns (Qiagen). Paired-end libraries were generated using the Illumina TruSeq RNA sample preparation kit. An Illumina HiSeq instrument was used for sequencing with paired-end read length of 100 bp.

Chromatin immunoprecipitation and sequencing. Chromatin immunoprecipitation was performed as described previously[39]. Briefly, 20,000–50,000 HSCs (SP^{KLS}CD150$^+$) were sorted and cross-linked with 1% formaldehyde at room temperature for 10 min, and the reaction was stopped by addition of 0.125 M glycine and incubation at room temperature for 5 min. Then, 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 of lysis buffer (10 mM Tris, pH 7.5, 1 mM EDTA, 1% SDS), samples were diluted with 150 μl of PBS with PIC and DNA was sonicated to generate fragments of 200–500 bp in length (Bioruptor, Diagenode). Sonicated chromatin was centrifuged at 4 °C for 5 min at 13,000 r.p.m. (15,871*g*) to remove precipitated SDS. A 180-μl aliquot was transferred to a fresh 0.5-ml collection tube, and 180 μl of 2× 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. For input control, 1/10 volume (36 μl) was removed. ChIP-qualified antibodies (0.1 μg of antibody to H3K4me3, Millipore, 07-473; 0.3 μg of antibody to H3K27me3, Millipore, 07-449) were added to the sonicated chromatin, and samples were incubated at 4 °C overnight. After this incubation, 10 μl of protein A magnetic beads (Dynal, Invitrogen) previously washed in RIPA buffer was added, and samples were incubated for an additional 2 h at 4 °C. Bead-protein complexes were washed three times with RIPA buffer and twice with TE buffer (10 mM Tris, pH 8.0, 1 mM EDTA). After transfer to a fresh 1.5-ml collection tube, genomic DNA was eluted for 2 h at 68 °C in 100 μl of 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 of elution buffer (20 mM Tris, pH 7.5, 5 mM EDTA, 50 mM NaCl) for 10 min at 68 °C. Precipitated DNA was purified by MinElute Purification kit (Qiagen) and eluted in 12 μl of elution buffer. Precipitated DNA was successfully used to generate a library with the ThruPLEX-FD preparation kit without additional PCR amplification (Rubicon). Sequencing was performed according to the manufacturer's protocol on a HiSeq 2000 instrument. 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 data. WGBS data analyses were based on BSMAP[40] and a newly developed program, MOABS (Model-Based Analysis of Bisulfite Sequencing) (D.S., Y. Xi, B.R., Y.J. Park, T. Pan *et al.*, unpublished data). We used four modules of MOABS—mMap, mCall, mOne and mComp. MOABS seamlessly integrates alignment, methylation ratio calling and the identification of hypomethylation for one sample and differential methylation for multiple samples, and this software also performs other downstream analysis.

Read mapping. BSMAP[40] was used to align paired-end reads from bisulfite-treated samples to the mm9 mouse genome. 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 could be mapped to multiple locations with the same minimal number of mismatches, this read was determined to be a multimapped read, and its mapping location was randomly selected from all mapping locations.

Quality control and methylation ratio calling. BSeQC[41] was used to remove technical biases in WGBS data. First, we removed clonal reads with identical sequences resulting from possible overamplification during sample preparation. These clonal reads were mapped to exactly the same position on the genome and could be identified on the basis of their extremely high coverage relative to the mean coverage across the genome, using a Poisson *P*-value cutoff of 1×10^{-5}. As a result, a maximum of two reads that mapped to the same location were kept for downstream analysis. Second, during adaptor ligation in bisulfite library preparation, the overhangs of DNA fragments were end repaired using unmethylated cytosines. This end-repair procedure might introduce artifacts if the repaired bases contained methylated cytosines. We modeled the overhang size of DNA fragments and determined that trimming three bases (the overhang size) from the repaired end was sufficient to eliminate nearly all artifacts introduced by end repair. Third, the overlapping segment of two read mates derived from the same DNA fragment was only processed once to prevent overcounting of the same DNA. Finally, the methylation ratio of each CpG was measured as the proportion of unconverted CpGs in all mapped reads, including both strands.

Differentially methylated regions. We used a first-order HMM to determine DMRs. For a two-sample comparison $p_2 - p_1$, the state of the *i*th CpG in the genome was denoted as S_i, where S_i can take three hidden states

$$\begin{aligned} &S_0\text{: hypomethylation state, if } p_2 - p_1 < -v_0 \\ &S_1\text{: state of no difference, if } |p_2 - p_1| < v_0 \\ &S_2\text{: hypermethylation state, if } p_2 - p_1 > v_0 \end{aligned}$$

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

For each CpG in the genome, we determined four numbers in total from two 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 CpGs in all mapped reads. Given our observations for all CpGs, we wanted to find the HMM that maximized the probability for each observation. The HMM is characterized by initial state π_0, transition probability matrix $A = \Pr(S_i \mid S_{i-1})$ and emission probability matrix $B = \Pr(x_i \mid S_i)$. The initial state π_0 can be assigned as S_1. By assuming that a cytosine is in one of the three hidden states, the emission probability for the *i*th CpG of $x = (n_1, k_1, n_2, k_2)$ when the state of the cytosine is S_i can be derived as follows:

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

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. As the training is computationally intensive, MOABS chooses only a subset of CpGs for the training, such as the first 1 million CpGs on chromosome 19 or CpGs provided by the user. After the change in likelihood for the model was smaller than a given threshold or the maximum number of iterations was reached, the optimal hidden state for each CpG was obtained. Consecutive CpGs with the same hypo- or hypermethylation state were merged as DMRs.

Undermethylated regions. As in detection of DMRs, we used a two-state first-order HMM to detect highly methylated and weakly methylated regions from a single sample. Only locations with coverage of more than ten reads were considered to increase the detection accuracy. Consecutive CpGs with the same hidden low-methylation state were merged to form an LMR. We also performed a random shuffle of all the CpGs in the genome, which was followed by the same procedure for LMR detection. The resulting null distribution indicates the number of CpGs required for LMR detection. With FDR set at 5%, each LMR will include at least four CpGs for wild-type HSCs or at least five CpGs for *Dnmt3a*-null HSCs. UMRs are a subset of LMRs with a 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 was still less than 10%. UMRs less than 1 kb long were not included in analysis in this manuscript. UMRs of 3.5 kb or longer were defined as canyons. UMRs 1 kb or more in length but less than 3.5 kb were used as cUMRs for comparison with canyons. See the MOABS website for further details.

Analysis of 5hmc CMS pulldown and histone modification ChIP-seq data. We sequenced 5hmC CMS samples as paired-end 100-bp reads. Reads were mapped to the mm9 mouse genome using BSMAP[40] by allowing at most four mismatches. Only uniquely mapped reads were used for MACS[42] peak calling with a *P*-value cutoff of 1×10^{-5}. Peaks were regions with enrichment in the CMS pulldown sample compared to the control sample. The control sample was sonicated, and bisulfite conversion was performed, but without CMS pulldown.

Common peaks were those that overlapped in wild-type and knockout samples, and sample-specific peaks were those that did 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, and a Poisson test was performed to detect whether one sample had more reads than the other sample at each synthetic peak. Before the test, the read number was normalized to 10 million for every sample.

The same pipeline was used to analyze histone modification ChIP-seq data, with a few exceptions. Histone modification ChIP-seq reads were mapped to the mm9 mouse genome using SOAP2 (ref. 43) by allowing at most two mismatches for 50-bp short reads and at most four mismatches for 100-bp longer short reads. Only uniquely mapped reads were kept. To remove PCR-generated duplicate reads, a maximum of two duplicate reads was allowed for each biological replicate. This maximum number was chosen on the basis of the 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. Uniquely mapped and duplicate removed reads from each biological replicate were fed as treatment files into the MACS program to find the enriched peaks. H3K4me3 peaks were called by MACS with default parameters, except that *P* values were set at 1×10^{-8}. H3K27me3 peaks were called by SICER with parameters 'window size 200 fragment size 200 gap size 600 and FDR 1E-8'. Peaks from all biological replicates of a specific sample were merged to form the final set of peaks for this specific sample.

Analysis of RNA sequencing data. Paired-end 100-bp reads were sequenced by RNA-seq. The last 20 bases were trimmed owing to average low quality. Alignment was performed by RUM[44], which first mapped reads to the genome and transcriptome by Bowtie and then used BLAT to remap to the genome reads that were initially unmapped. Information from the two rounds of mapping was merged, and multiply mapped reads were discarded. Gene annotations used for transcriptome alignment included RefSeq, UCSC knownGene and Ensembl gene models. Gene expression (in FPKM) was measured by counting the reads matching the exons of each gene. Differential expression was analyzed using edgeR[45].

UMR dynamics in size and methylation ratio. We defined UMR dynamics in size, including expansion, contraction and no change, between wild-type and *Dnmt3a*-null samples using the following criterion: if one edge of a wild-type UMR moved outward or inward in the *Dnmt3a*-null sample for more than 200 bases, this edge was classified as expanded or contracted, respectively. If the change involved fewer than 200 bases, the edge was classified as unchanged. Furthermore, if the wild-type UMR disappeared in the *Dnmt3a*-null sample, both edges of the UMR were classified as contracted; in contrast, both edges of an emerging UMR in the *Dnmt3a*-null sample were classified as expanded.

We merged all UMRs in both wild-type and *Dnmt3a*-null samples into 19,569 synthetic UMRs and measured the contribution of each sample to the length of each synthetic UMR. Sample-specific contribution to a given synthetic UMR was defined as the length of the sample-specific UMR divided by the length of the synthetic UMR. The contribution from the *Dnmt3a*-null sample for almost all synthetic UMRs was close to 1, indicating global UMR expansion in the *Dnmt3a*-null sample. Strikingly, 16% of synthetic UMRs emerged in the *Dnmt3a*-null sample and thus had no contribution at all from the wild-type sample. In contrast, only 4% of synthetic UMRs disappeared completely in the *Dnmt3a*-null sample. Furthermore, each of the 508 synthetic UMRs had multiple wild-type UMRs separated by methylated CpGs that were eroded in the *Dnmt3a*-null sample, such that these multiple UMRs were connected to a longer UMR. In contrast, only 177 wild-type UMRs were broken down into multiple UMRs in the *Dnmt3a*-null sample.

To test whether a given synthetic UMR was differentially methylated, we compared the mean methylation ratio of the synthetic UMR in wild-type and *Dnmt3a*-null samples using MOABS, with the permutation FDR set at 0.2%. The results indicated that 14% of synthetic LUMRs were differentially methylated in wild-type and *Dnmt3a*-null samples.

Analysis of Oncomine AML genes. We used Oncomine (Compendia Biosciences) to assess the enrichment of canyon-associated genes expressed in wild-type mouse HSCs (FPKM > 1) in signatures of genes overexpressed in leukemic disease versus normal bone marrow. Oncomine assesses overlap significance with Fisher's exact test. Our threshold criteria were an odds ratio of ≥1.8 and a P value of $<1 \times 10^{-5}$. To address the challenge of cross-species comparison as well as the inherent technical limitations of comparing next-generation sequencing data to those derived from legacy microarray technologies, we limited our analysis to signatures derived from the 2 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[46]. Genes from simulated canyons represent randomly sampled genes with promoter enrichment or depletion of H3K27me3 and/or H3K4me3 histone modifications proportionate to the distributions observed in wild-type HSC canyons, as determined by ChIP-seq (**Fig 2a**). Gene set distribution was as follows: H3K4me3$^+$H3K27me3$^+$ (45.92%), H3K4me3$^-$H3K27me3$^+$ (6.97%), H3K4me3$^+$H3K27me3$^-$ (46.56%) and H3K4me3$^-$H3K27me3$^-$ (0.54%). Bivalent promoters (H3K4me3$^+$H3K27me3$^+$) required an overlap of at least one nucleosome length (~146 bp) between H3K4me3 and H3K27me3 peaks. Random unmethylated promoters were sampled from promoters (excluding canyon-associated genes) with a mean CpG methylation level of <10% in wild-type HSCs. Promoter regions were defined as 2-kb regions centered on the transcription start site in RefSeq transcripts. Random expressed genes were sampled from genes with FPKM of >1 in wild-type HSCs. Details on all gene sets and Oncomine signatures represented in the analysis are provided in **Supplementary Table 8**.

Analysis of TCGA AML genes. We downloaded RNA-seq data for individuals with AML from the TCGA Data Portal (see URLs) and performed preprocessing; $\log_2$ transformation, orthologous gene mapping and filtering out of genes with over 20% missing values. In the process, human gene symbols were mapped to those for mice using human-mouse homology information from Mouse Genome Informatics (see URLs). Finally, we selected the gene expression data for 14,701 genes in 167 cases. Two-sample t tests were applied to identify significantly differentially expressed genes between two groups based on *Dnmt3a* mutation status. We selected 1,760 signature genes ($P < 0.05$). BRB-ArrayTools and the R language (see URLs) were primarily used for statistical analysis of gene expression data[47]. Cluster analysis was performed using Cluster software, and heat maps were generated with Treeview[48]. We assessed the enrichment of expressed canyon-associated genes (FPKM > 1) in *Dnmt3a* mutation signatures using the hypergeometric test in the R language.

Analysis of canyon-associated genes in CCDE. Gene expression in 947 cancer cell lines from CCDE (GSE36139) was used for hierarchical clustering of canyon-associated genes. Cluster analysis was performed using Cluster software, and heat maps were generated with Treeview[48].

34. Goodell, M.A., Brose, K., Paradis, G., Conner, A.S. & Mulligan, R.C. Isolation and functional properties of murine hematopoietic stem cells that are replicating *in vivo. J. Exp. Med.* **183**, 1797–1806 (1996).
35. Mayle, A., Luo, M., Jeong, M. & Goodell, M.A. Flow cytometry analysis of murine hematopoietic stem cells. *Cytometry A* **83**, 27–37 (2013).
36. Gu, H. *et al.* Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat. Protoc.* **6**, 468–481 (2011).
37. Huang, Y., Pastor, W.A., Zepeda-Martinez, J.A. & Rao, A. The anti-CMS technique for genome-wide mapping of 5-hydroxymethylcytosine. *Nat. Protoc.* **7**, 1897–1908 (2012).
38. Krueger, F. & Andrews, S.R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
39. Dahl, J.A. & Collas, P. A rapid micro chromatin immunoprecipitation assay (microChIP). *Nat. Protoc.* **3**, 1032–1045 (2008).
40. Xi, Y. & Li, W. BSMAP: whole genome Bisulfite Sequence MAPping program. *BMC Bioinformatics* **10**, 232 (2009).
41. Lin, X. *et al.* BSeQC: quality control of bisulfite sequencing experiments. *Bioinformatics* doi:10.1093/bioinformatics/btt548 (11 October 2013).
42. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
43. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
44. He, Y.-F. *et al.* Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303–1307 (2011).
45. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
46. Kuhn, A., Luthi-Carter, R. & Delorenzi, M. Cross-species and cross-platform gene expression studies with the Bioconductor-compliant R package 'annotationTools'. *BMC Bioinformatics* **9**, 26 (2008).
47. Simon, R. *et al.* Analysis of gene expression data using BRB-ArrayTools. *Cancer Inform.* **3**, 11–17 (2007).
48. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).