Supplementary information

Large conserved domains of low DNA methylation maintained by Dnmt3a

Mira Jeong#1, Deqiang Sun #2, Min Luo#1, Yun Huang³, Grant A. Challen%1, Benjamin Rodriguez², Xiaotian Zhang¹, Lukas Chavez³, Hui Wang⁴, Rebecca Hannah⁵, Sang-Bae Kim⁶, Liubin Yang¹, Myunggon Ko³, Rui Chen¹, Berthold Götgens⁶, Ju-Seog Lee⁶, Preethi Gunaratne⁷, Lucy A. Godley⁸, Gretchen J. Darlington⁹, Anjana Rao³, Wei Li², and Margaret A. Goodell¹

# These authors contributed equally to this work
% These authors jointly directed this work

Affiliations:
1Stem Cells and Regenerative Medicine Center, Department of Pediatrics and Molecular & Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA.
2Division of Biostatistics, Dan L. Duncan Cancer Center and Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, Texas 77030, USA.
3Division of Signaling and Gene Expression, La Jolla Institute for Allergy and Immunology, La Jolla, California 92037, USA
4Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA.
5Department of Hematology, Cambridge Institute for Medical Research and Wellcome Trust and MRC Cambridge Stem Cell Institute, Cambridge University, Hills Road, Cambridge, UK.
6Department of Systems Biology, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX 77054, USA.
7Department of Pathology, Baylor College of Medicine, and Department of Biology & Biochemistry, University of Houston, Houston, TX 77204, USA
8Department of Medicine, The University of Chicago, Chicago, Illinois 60637, USA.
9Huffington Center for Aging, Baylor College of Medicine, Houston, Texas 77030, USA.
%Now at the Department of Internal Medicine, Washington University in St. Louis, St. Louis, Missouri 63110, USA.

Correspondence to: Goodell@bcm.edu or WL1@bcm.edu
Supplementary Tables

Supplementary Table 1: Whole genome sequencing statistics
Supplementary Table 2: UMRs in WT HSCs
Supplementary Table 3: UMRs in HOX genes
Supplementary Table 4: UMRs in ESCs
Supplementary Table 5: Canyon vs UMRs
Supplementary Table 6: WT vs Dnmt3a KO HSC UMRs
Supplementary Table 7: CMS 5hmC sites in WT and Dnmt3a KO
Supplementary Table 8: oxBS sequencing
Supplementary Table 9: Summary of genes overlapping leukemia Oncomine signatures
Supplementary Figure 1. Genomic features of the WT HSC methylome. Violin plots indicate the methylation ratios in regions with different genomic features.
Supplementary Figure 2. Canyon distribution in mouse chromosomes. 1,104 Canyons are plotted along different chromosomes.
Supplementary Figure 3. Large unmethylated Canyons around Hox genes. (a) Bar graph indicating the percentage of genes overlapping Hox genes for Canyons, other UMRs and control genes. Control genes are the merged version of three control gene sets: other transcription factors (TFs), other CGI and Bystander genes. The Y-axis indicates percent of overlapping genes. (b) Length distribution and Homeobox gene enrichment of the largest 20 canyons. Green star indicates Canyon genes expressed in HSCs. (c) UCSC genome browser track depicts methylation profile across the Barhl2 gene, (d) Uncx gene and (e) Pou3f2 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 indicated in dark blue. RNA-seq expression is shown at bottom in green.
Supplementary Figure 4. Mammalian conservation scores and depletion of transposable elements in Canyons (a) Bar graph indicating the average phast conservation scores for Canyons, other UMRs and control genes. Control genes are the merge of three control gene sets: other TFs, other CGI and Bystander genes. (b) Position of repeats, including LINE, LTR and SINE on WT HSC Canyons. The X-axis of the plots indicates location of repeats relative to Canyons, where all Canyons are normalized to cover positions -1 to +1. Thus, a position weight of 0 represents the middle of the Canyons, while position weight ±1 represents the edges of the Canyons. (c) Bar graph indicating the mean percentage of the indicated regions that are covered by repeats.
Supplementary Figure 5. HSC-specific TF binding peak enrichment in Canyons. (a) UCSC genome browser track depicts overlap of TF binding peaks and Canyons, using over 150 ChIP-seq data sets including 10 hematopoietic stem cell TFs (Scl/Tal1, Lyl1, Lmo2, Gata2, Runx1, Meis1, Pu.1, Erg, Fli1, and Gfi1b). Purple boxes indicate TF binding peaks, the green bar indicates a Super enhancer in T helper (Th) cells (Super enhancer data from ref3). (b) Position of binding peaks for the 10 hematopoietic stem cell TFs across all UMRs. ±1 represents the boundary of cUMRs. (c) Bar graph indicating the mean percentage covered by TF binding peaks for Canyons and other UMRs and control genes. Control genes are the merge of three control gene sets: other TFs, other CGI and Bystanders1 for negative controls.
Supplementary Figure 6. Canyon dynamics between ESC and HSC. (a) Dot plot to compare Canyon methylation ratios between HSCs and ESCs. Conserved Canyons are boxed in green, and cell-type-specific Canyons are boxed in red and blue with examples from each group depicted in b-e. (b and c) The UCSC browser image shows DNA methylation profiles for cell type-specific Canyons Pou5f1 (Oct4) (ESC-specific) and Erg (HSC-specific). (d and e) Depiction of the common Canyons Brd2 and Lhx5 genes. Lhx5 is not expressed in either cell type, while Brd2 is only expressed in HSCs (RNA seq data not shown). (f) Stable Pax7 Canyon in human hematopoietic progenitors and differentiated progeny (data from ref3). (g) Stable Pax7 Canyon in mouse HSC and ESC (ES data from ref4). (h) Bar graph to compare common canyon numbers between different cell types and different species. (Mouse ES data from ref4), Human HPC, B cell and Neutrophil data from ref3, and other human cell data from NIH Roadmap Epigenomics Mapping Consortium (www.roadmapepigenomics.org). HumanCD34: Mobilized CD34+ primary cells, Human Liver: Adult Liver, Human Brain: Brain Germinal Matrix, Human neurospere: Neurosphere cultured cells—Ganglionic Eminence Derived. Y-axis indicates common Canyon numbers and percentages.
Supplementary Figure 7. Dnmt3a is involved in maintaining Canyons in HSCs. (a) Position of differentially methylated regions (DMRs) comparing WT and Dnmt3a KO HSCs on cUMRs and Canyons. The grey bars represent controls, generated by assigning the UMRs and Canyons to random locations in the genome. The X-axis indicates the genomic position centered on the midpoint of UMRs normalized to -1 to +1. Thus, ±1 represents the boundaries of UMRs and Canyons. (b) UCSC genome browser track depicts methylation profiles across HoxB cluster. Several smaller Canyons merge and expand in the Dnmt3a KO HSCs to generate exceptionally large under-methylated regions.
Supplementary Figure 8. 5hmC peak enrichment on Canyon boundaries. (a) UCSC genome browser track depicts FAM32a gene which is one of the genes covered by at least 100 reads in BS-seq and oxBS-seq. Y-axis show % of 5mC or % of 5hmC. +/- % range indicates plus strand or minus strand. (b) Ccdc124 gene. (c) Position of CMS peaks in WT HSC on normalized cUMRs and Canyons. The grey bars represent the control generated by assigning the UMRs and Canyons to random locations in the genome. The X-axis of the plots indicates genomic positions centered on the midpoint of UMRs, with ±1 representing the boundaries of the UMRs.
Supplementary Figure 9. Alterations of 5hmC peaks in the hypomethylated DMRs after Dnmt3a KO. (a) UCSC genome browser track depicts methylation profiles and 5hmC, detected by the CMS method, across the Srrm2 gene in WT and Dnmt3a KO HSCs. Blue box indicates methylation depleted region with decreased 5hmC signal and pink box indicates slightly methylation decreased region with increased 5hmC signal. (b) UCSC genome browser track depicts methylation profiles and CMS peaks across the Hoxb4 gene in WT and Dnmt3a KO HSCs. Blue box indicates methylation depleted region with decreased 5hmC signal and pink box indicates slightly methylation decreased region with increased 5hmC signal. (c) When the methylation is depleted in Dnmt3a KO (MethylationRatio<20%), the majority of 5hmC peaks showed decreased density. (d) When the methylation decreases but still retains a certain level in Dnmt3a KO (40%<MethylationRatio<80%), 3-fold more 5hmC peaks showed increased density. Each dot represents a CMS peak, which overlaps with hypomethylated DMR. X-axis and Y-axis values are normalized CMS read counts in the peak region for WT and Dnmt3a KO HSC. Grey dots represent those peaks that are not significantly changed, while red and blue dots represent those peaks that are significantly increased and decreased.
Supplementary Figure 10. Unsupervised clustering analysis of Canyon genes separates WT and DNMT3A mutant in 173 TCGA AML. Gene expression data from 173 TCGA AML patients by RNA-seq were used for hierarchical clustering of Canyon genes. Expression data were centralized by subtracting median expression level across samples. The data are presented in matrix format in which rows represent individual genes and columns represent each patient. Relatively increased expression is indicated by red color; relatively decreased expression by green color; intermediate expression by black color as indicated in the scale bar (log2 transformed scale). UC: Unexpressed Canyon genes (FPKM < 1) EC: Expressed Canyon genes (FPKM ≥ 1).
Supplementary Figure 11. Heatmap of unsupervised clustering analysis of Canyon gene expression from 947 Cancer Cell Line Encyclopedia (CCLE). Gene expression data from 947 cancer cells (GSE36139) by microarray were used for hierarchical clustering of Canyon genes. Expression data were centralized by subtracting median expression level across samples. The data are presented in matrix format in which rows represent individual genes and columns represent each cancer cell line. Relatively increased expression is indicated by red color; relatively decreased expression by green color; intermediate expression by black color as indicated in the scale bar (log2 transformed scale). UC: unexpressed Canyon genes (FPKM < 1) EC: Expressed Canyon genes (FPKM ≥ 1).
Supplementary Figure 12. Model to show the proposed mechanism of Canyon size dynamics. Dnmt3a and Tet proteins act at the edges of Canyons. As long as both proteins are present, Canyon size is maintained in successive rounds of cell division. Canyon expansion is more frequently associated with high levels of H3K4me3 and the presence of expressed genes. In the absence of Dnmt3a, the presence of 5hmC (deposited by Tet proteins) results in edge erosion. In the absence of Tet proteins, the absence of deposition of 5-hydroxymethylation is predicted to result in Canyon contraction.
References for Supplementary Data