The Competition between DNA Methylation and Demethylation is Associated with Transcription Regulation and Tumorigenesis

Wei Li (wei.li@uci.edu)
University of California, Irvine https://orcid.org/0000-0001-9931-5990
Jiejun Shi
University of California, Irvine
Jianfeng Xu
Baylor College of Medicine
Yiling Chen
University of California, Los Angeles
Jingyi Jessica Li
University of California Los Angeles https://orcid.org/0000-0002-9288-5648
Lanlan Shen
Baylor College of Medicine

Article

Keywords: DNA methylome, DNA methyltransferases, demethylation, competition effect, gene expression, transcription regulation, tumorigenesis

Posted Date: January 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-123209/v1

Version of Record: A version of this preprint was published at Nature Communications on September 6th, 2021. See the published version at https://doi.org/10.1038/s41467-021-25521-7.

The Competition between DNA Methylation and Demethylation is Associated with Transcription Regulation and Tumorigenesis

Jiejun Shi[1], Jianfeng Xu[2], Yiling Elaine Chen[3], Jingyi Jessica Li[3], Lanlan Shen[4], Wei Li[1*]

Affiliations:

[1]Division of Computational Biomedicine, Department of Biological Chemistry, School of Medicine, University of California, Irvine, Irvine, CA 92697, USA

[2]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX 77030, USA

[3]Department of Statistics, University of California, Los Angeles, CA 90095, USA

[4]Department of Pediatrics, Baylor College of Medicine, USDA/ARS Children's Nutrition Research Center, Houston, Texas, 77030, USA

*Correspondence: wei.li@uci.edu

Abstract

The mammalian DNA methylome is formed by two antagonizing processes, methylation by DNA methyltransferases (DNMT) and demethylation by ten-eleven translocation (TET) dioxygenases. Although the dynamics of either methylation or demethylation have been intensively studied in the past decade, their competition effect remains elusive. Here, we quantify the competition between DNA methylation and demethylation by the percentage of unmethylated CpGs within a partially methylated read from bisulfite sequencing. After verifying methylation competition by its strong association with the co-localization of DNMT and TET enzymes, we observe that methylation competition is strongly correlated with gene expression. In particular, during tumorigenesis, the elevation of methylation competition is associated with the repression of 40~60% of tumor suppressor genes, which cannot be explained by promoter hypermethylation alone. Furthermore, methylation competition can be used to stratify large undermethylated regions with negligible differences in average methylation into two subgroups with distinct chromatin accessibility and gene regulation patterns. Together, methylation competition represents a novel methylation metric important for transcription regulation and tumorigenesis and is largely distinct from conventional metrics, such as average methylation and methylation variation.

Introduction

DNA methylation at CpG dinucleotide(5mC) is introduced and maintained by DNA methyltransferases (DNMT family)[1,2]. Meanwhile, through hydroxymethylation, 5mC is removed by ten-eleven translocation dioxygenases (TET family)[3,4]. Besides their opposing effects, the two enzyme families present complementary DNA binding patterns. While TET1 protein prevents *de novo* methyltransferases from binding to regulatory elements[5,6], DNMT3A also blocks TET1 binding, especially in promoter regions[6]. Interestingly, these two 'competing' enzyme families are observed to be jointly associated with tumor malignancy. For example, the *DNMT3A* and *TET2* double-knockout mice show worse survival than single-knockout counterparts[7]; also, mutations in *DNMT3A* and *TET2* significantly co-occur in human T-cell lymphoma[8]. These findings suggest that the competition between methylation and demethylation processes is related to tumorigenesis. However, to what extent this competition contribute to cancer gene regulation remains largely unknown.

For years, DNA methylation levels have been quantified in an 'average' manner. The increased average methylation level of CpG island (CGI), i.e., CGI hypermethylation, is a well-established mechanism for gene silencing[9]. Numerous differentially methylated regions have been identified based on between-sample comparison of average methylation levels[10,11]. Besides the average methylation, DNA methylation has been quantified by its variation as 'methylation heterogeneity'[12] or 'epigenetic polymorphism'[13]. Methylation heterogeneity scores are defined based on the frequencies of methylation patterns (epialleles) at multiple CpGs inferred from

bisulfite sequencing reads[12–15]. Previous studies reveal that methylation variation is associated with global transcription variation[12,16,17]. However, neither average methylation nor methylation variation can delineate the degree of competition between active methylation and demethylation.

Recently, a mathematical model has been applied to deconvolute methylation and demethylation rates from average methylation levels of individual CpGs in stem cells[18]. However, the spatial coupling of methylation competition at adjacent CpGs has not been considered, and such coupling may be critical for transcription factor (TF) binding and cancer gene regulation[19].

Bisulfite sequencing has enabled the measurement of DNA methylation of adjacent CpGs within the same read[20], and thus it captures methylation competition if there are unmethylated CpGs in a partially methylated read. Through integrative analysis of 75 methylomes and 44 transcriptomes, we demonstrate that methylation competition unveils a new type of methylation abnormality, which is largely distinct from both the change of average methylation and methylation variation. We find that methylation competition is associated with a previously undetected repertoire of epigenetically regulated tumor suppressor genes (TSGs), and it can be used to stratify large undermethylated regions into two subgroups with distinct characteristics in chromatin accessibility and gene regulation.

Results

Delineating the competition between active DNA methylation and demethylation

We quantify the competition between active DNA methylation and demethylation within the same cell by dissecting reads from bisulfite sequencing. The active competition events are captured by the unmethylated CpG(s) within a partially methylated read (slashed circles in Fig 1a) because each read comes from one cell. Fully methylated and unmethylated reads, in contrast, do not possess active competition information, as they are dominated by methylation and demethylation, respectively. Hence, we dissect bisulfite sequencing reads into three categories of fragments (or sub-reads), i.e., methylated fragments (with each fragment consisting of consecutive solid circles in Fig 1a, denoted as '*M'*), unmethylated fragments (with each fragment consisting of consecutive blank circles, denoted as '*U'*), and methylation competition fragments (with each fragment consisting of consecutive slashed circles, denoted as '*C'*). Although both '*U'* and '*C'* fragments are unmethylated, they are different: *'U'* fragments are the fully unmethylated reads, while *'C'* fragments are segments of unmethylated CpGs within partially methylated reads. We define the methylation competition ratio of a genomic region as the sum of '*C'* fragments' weights divided by the sum of all fragments' weights in that region (Eq.1 in Methods). Each fragment's weight is set as its number of CpGs (see Methods).

We compare the methylation competition ratio with two measures of the average methylation (i.e., the traditional mean methylation and the cellular heterogeneity-adjusted clonal methylation (CHALM)[21], see Methods) and three measures of the methylation variation (i.e., Shannon's entropy[15], Epipolymorphism[13], and the proportion of discordant reads (PDR)[12], see Methods) across gene promoters using whole-genome bisulfite sequencing (WGBS) data from mouse embryonic stem cells (mESCs). We observe that these six methylation measures are correlated to different degrees with DNA binding intensities of DNMT3A1 and TET1 enzymes in matched sample measured by ChIP-seq (chromatin immunoprecipitation followed by high-throughput sequencing). We observe that the average methylation measures are correlated positively and negatively with the binding intensities of the methyltransferase DNMT3A1 and the demethylase TET1, respectively, consistent with the known enzymatic activities of the two enzymes (Supplementary Fig 1b and 1c). The methylation variation measures show similar correlation patterns (Supplementary Fig 1d, 1e, and 1f). In contrast, the methylation competition ratio is positively correlated with both DNMT3A1 and TET1 binding intensities (Supplementary Fig 1a). To further explore the two enzymes' joint effects, we define the DNMT3A1-TET1 'joint regulation score' (Π) of a promoter as the product of DNMT3A1 and TET1 binding intensities within that promoter (see Eq.2 in Methods). This joint regulation score depicts the extent to which the promoter is co-occupied by both enzymes, and it takes a low value if either enzyme has a low binding intensity. As expected, the joint regulation score has a strong positive correlation with the

methylation competition ratio (Fig 1b) but not as much with the average methylation measures (Fig 1c and Supplementary Fig 1c) or the methylation variation measures (Fig 1d and Supplementary Fig 1e and 1f). This result is consistent with the fact that average methylation and methylation variation are neutralized by the additive effects of two opposing enzymes and that only the methylation competition ratio characterizes the competition between DNMT3A1 and TET1. In the following text, we use the traditional mean methylation as the average methylation measure and Shannon's entropy as the methylation variation measure.

Based on DNMT3A1 and TET1 co-localization patterns, we categorize promoters into four groups: DNMT3A+TET1+, DNMT3A+TET1-, DNMT3A-TET1+, and DNMT3A-TET1-, where + and – indicate strong and weak bindings, respectively. As shown in Fig 1e, the methylation competition ratio is significantly higher in 'DNMT3A+TET1+' promoters than in the other three groups of promoters, which are bound by only one or none of the two enzymes. For example, *Prok2* (Fig 1f) has a DNMT3A+TET1+ promoter and a high methylation competition ratio, while *Ogt* (Fig 1g) has a DNMT3A-TET1- promoter and a low methylation competition ratio. Notably, only the methylation competition ratio can distinguish the different co-localization patterns of *Prok2* and *Ogt*, while the average methylation and the methylation variation cannot. Moreover, only the methylation competition ratio is predictive of the expression levels of *Prok2* and *Ogt* (*Prok2* has a high ratio and low expression, while *Ogt* has a low ratio and high expression), while the average methylation and the methylation variation are not (both genes have similar average

methylation and methylation variation levels). These findings are supported by external evidence in mouse samples with *Dnmt* or *Tet* knockout. As expected, *Dnmt3a* knockout leads to genome-wide hypomethylation, whereas *Tet* knockout leads to global hypermethylation (Fig 1h). We observe that either knockout results in more regions with decreased methylation competition ratios than regions with increased ratios, an observation consistent with our definition of methylation competition ratio (Fig 1i). Overall, these results confirm that the methylation competition ratio, a new quantitative measure of methylation, can delineate the antagonism between methylation and demethylation processes.

Methylation competition is negatively correlated with gene expression

Our previous analysis of genes *Prok2* and *Ogt* (Fig 1f and 1g) suggests that the methylation competition ratio may be a better predictor of gene expression than the average methylation and the methylation variation. To further examine the relationship between these quantitative measures of methylation and gene expression, we first calculate the methylation competition ratios, the average methylation, and the methylation variation of every CpG site and transcription regulatory elements/regions of three types (promoters, gene-body regions, and enhancers, see Methods) using WGBS data of primary cells and normal tissues from Epigenomic Roadmap Consortium[22]. Then we quantify gene expression levels using the RNA sequencing (RNA-seq) data from matched samples (Supplementary Table 1) and separate all genes into four equal-sized groups based on expression quantiles (0-25%, 25-50%, 50-75%, and 75-100%). As indicated in Fig 2a and

Supplementary Fig 2a, more highly expressed genes are characterized by lower methylation competition ratios, lower average methylation (by both the traditional mean and CHALM), and lower methylation variation in promoters. However, unlike the average methylation and the methylation variation, the methylation competition ratios in TSS-proximal regions (<2kb) are higher than in more distal regions (>2kb), suggesting that the antagonism between methylation and demethylation is more intense near transcription start sites. Our findings are in line with a previous knockout study, which reveals that DNMT3A and TET1 prevent binding of each other mainly in TSS-proximal regions[6].

Compared with the average methylation (measured by both the traditional mean and CHALM) and the methylation variation, the methylation competition ratio is better correlated with gene expression, and this phenomenon is consistent across all three types of regulatory elements in CD3+ T-cells (Fig 2b and Supplementary Fig 2b). In contrast, the correlations between gene expression and the average methylation or the methylation variation are not consistent across the three types of regulatory elements; for example, CHALM values in gene body regions and enhancers have much lower correlations with gene expression than CHALM values in promoters do (Supplementary Fig 2b). We further confirm this phenomenon using WGBS and RNA-seq data of other primary cells, fetal tissues, and adult tissues (Supplementary Fig 3). All together, we show that the methylation competition ratio is a better predictor of gene expression than both the average methylation and the methylation variation.

Effects of sequencing depth and CpG density on quantification of methylation competition

Sequencing depth is a key factor that affects quantitative analysis of high-throughput sequencing data. Using a down-sampling analysis, we observe an expected phenomenon that the methylation competition ratio has a stronger negative correlation with gene expression when sequencing depth increases, and the correlation plateaus at ~60x depth (Supplementary Fig 5). Notably, across sequencing depths from ~4x to ~86x, the methylation competition ratio consistently has better correlations with gene expression than the average methylation and the methylation variation do.

Besides sequencing depth, CpG density also affects the correlations between methylation measures and gene expression. To investigate this issue, we stratify genes into three groups based on the CpG densities in their promoters, i.e., high-CpG promoter (HCP) genes, intermediate-CpG promoter (ICP) genes, and low-CpG promoter (LCP) genes[23]. For all three methylation measures (the methylation competition ratio, the average methylation, and the methylation variation), the correlations between their values in promoters and gene expression decrease from HCP genes to ICP genes and further down to LCP genes. Among the three measures, the methylation competition ratio consistently has the strongest negative correlation with gene expression for all three groups of genes (Supplementary Fig 6).

Methylation competition is associated with the repression of tumor suppressor genes

To investigate whether methylation competition is involved in tumorigenesis, we apply gene set enrichment analysis (GSEA) to profiles of methylation competition ratios in promoters in 8 TCGA normal samples. As shown in Supplementary Fig 7a and 7b, the curated tumor suppressor genes (TSGs) in the COSMIC Cancer Gene Census (CGC) database[24] are associated with low methylation competition ratios. In contrast, cell lineage genes such as Homeobox genes are associated with high methylation competition ratios. In addition, housekeeping functions, e.g., 'pentose phosphate pathway', do not exhibit associations with methylation competition. Further analysis of 56 normal samples' methylomes shows a conserved pattern of low methylation competition at the promoter of *TP53*, a well-known TSG (Supplementary Fig 7c), and high methylation competition at the promoter of *NKX1-1*, a Homeobox transcription factor related to organ development and regeneration[25] (Supplementary Fig 7d).

Promoter hypermethylation is a well-established mechanism for TSG silencing in tumor[26]. We compare the methylation competition ratios and the average methylation between the normal and tumor uterus samples from the same patient in TCGA. Through an overlap analysis of genes that exhibit changes in methylation competition and/or average methylation, we identify three gene groups: 863 methylation-competition elevated and average-methylation stable genes (P1), 1,130 methylation-competition stable and average-methylation increased

(hypermethylated) genes (P2), and 482 methylation-competition elevated and average-methylation increased (hypermethylated) genes (P3) (Fig 3a). For the three gene groups, the average profiles of the methylation competition ratios and the average methylation in TSS-proximal regions are illustrated in Fig 3b. Compared with randomly selected genes, all three groups of genes demonstrate more significant transcription repression (i.e., smaller p-values) in the tumor uterus sample compared with the matched normal uterus sample by the Wilcoxon signed-rank test (Fig 3c). Through the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis, we find all three groups of genes enriched in various cancer-related signaling pathways (Fig 3d). We also find 27 TSGs in the three gene groups, and more than 44% (12/27) of them cannot be explained by promoter hypermethylation alone (P1 TSGs in Fig 3e). For example, *ZBTB16* (also known as PLZF) is a P1 TSG that inhibits prostate cancer tumor growth through the interplay with *PTEN* and *FOXO3a*[27], and its genetic alterations have been found in metastatic prostate cancer samples[28]. Although *ZBTB16* has an unclear function for uterine cancer, our result suggested that the elevation of methylation competition in its promoter may indicate a novel mechanism for its tumor suppressing role.

We also analyze the methylomes of normal and tumor breast samples from the same patient in TCGA (Supplementary Fig 8). Again, 66% (10/15) of repressed TSGs can be explained by the methylation competition elevation but not by hypermethylation, and *ZBTB16* is again an example (Supplementary Fig 8e).

To compare the methylation competition ratio again with the methylation variation (i.e., the methylation entropy), we perform a similar overlap analysis of genes that exhibit changes in methylation entropy and/or average methylation. We find that the 243 entropy elevated and average-methylation stable genes (Q1, Supplementary Fig 9a and 9b) are not significantly repressed in the tumor uterus sample than in the matched normal sample, and that they do not include any TSGs (Supplementary Fig 9c, 9d, and 9e). This result shows that, unlike methylation competition, methylation entropy is not associated with the repression of TSGs.

Collectively, our analysis suggests that the methylation competition reveals a repertoire of epigenetically regulated tumor suppressor genes that cannot be detected by the average methylation or the methylation variation.

Methylation competition in large undermethylated regions

Undermethylated regions (UMRs) are the clusters of adjacent, lowly methylated CpGs spanning from hundreds to thousands of base pairs[29]. Large UMRs, which are also termed as methylation 'canyons'[29] or 'valleys'[30], are conserved across cell types[29] and involved in chromatin interactions[31], and their target genes are associated with cell lineages[30]. Although all the methylation canyons are poorly methylated (with the average methylation ≤ 0.1), their methylation competition ratios range from 0 to 0.4 (Supplementary Fig 10d). To examine the role of methylation competition in their function, we classify the methylation canyons into high-competition canyons (blue in Fig 4) and low-competition canyons (red in Fig 4),

based on a cutoff derived from the genome background (see Supplementary Fig 10 and Methods). Although the two groups do not differ in their average methylation, their chromatin accessibilities are dramatically distinct (Fig 4a). Accordingly, we refer to the low-competition canyons, which are enriched with active markers (H3K4me3 and DNase I hypersensitive sites), as active canyons (aCanyons); we refer to the high-competition canyons, which are bound by H3K27me3, as Polycomb canyons (pCanyons).

Although the methylation variation measures (entropy, Epipolymorphism, and PDR) also differ between aCanyons and pCanyons (Supplementary Fig 11), they do not provide the same information as the methylation competition ratio does: only the methylation competition ratio is higher inside pCanyon than in the flanking regions. This phenomenon is consistent with the finding of a previous study that TET1 binding is elevated in canyons in *Dnmt3a* knockout sample[6].

Also, the TF binding patterns are consistent with the definitions of aCanyons and pCanyons. The aCanyons are bound by TFs involved in active transcription, such as transcription initiation factor TFIID subunit 1 (TAF1), TATA-box binding protein (TBP), and RBBP5, a subunit of MLL complex. In contrast, the pCanyons are bound by subunits of the PRC2 complex, such as SUZ12 and EZH2 (Supplementary Fig 12).

Recently, Zhang *et al.* reveal that methylation canyons are involved in chromatin loops that rely on Polycomb binding instead of cohesion or CTCF[31]. Consistent with this finding, we find that pCanyons are more enriched with chromatin interactions than aCanyons are (Supplementary Fig 13).

To gain further insights into the functions of canyons, we define canyon targets as the genes whose promoters or gene-body regions overlap with canyons. Again, consistent with the definitions of aCanyons and pCanyons, aCanyon targets are highly expressed, while pCanyon targets are almost silenced (Fig 4b). Gene ontology analysis reveals that aCanyon targets are enriched with TSGs and cancer pathways, while pCanyon targets are enriched with cell fate commitment and Homeobox genes. As a negative control, randomly selected genes are not enriched with these functional terms (Fig 4c).

Furthermore, we correlate the expression of canyon targets with the average methylation in a locus-specific manner[32]. Specifically, we first divide the promoter and downstream region (-2 to +10kb) of each gene into 120 equal-length bins. Then we compute the correlation between gene expression and the average methylation in each bin across aCanyon targets, pCanyon targets, and randomly selected genes. As expected, for all three gene sets, the average methylation in promoters is negatively correlated with gene expression (Fig 4d). Previous studies report positive correlations between the average methylation in gene-body regions and gene expression [10,33]. However, we observe that this phenomenon only occurs for

pCanyon targets and randomly selected genes, but not for aCaynon targets (Fig 4d). The surprisingly negative correlations we observe (between the average methylation in gene-body regions and gene expression) for aCanyon targets suggest that these genes may have a distinct methylation regulation mechanism.

Using *in vitro* SELEX assays, Yin *et al.* profile human TFs for binding preferences towards 5mC[34]. They classify the TFs into 'methyl-plus' and 'methyl-minus' groups, so that methyl-plus TFs' binding is enhanced by 5mC, while methyl-minus TFs' binding is inhibited by 5mC. To further understand aCanyons and pCanyons, we utilize this SELEX data and perform differential analysis of TF motif enrichment. We find that the TF motifs' 5mC preferences are negatively correlated with their fold enrichment between aCanyon and pCanyon (Fig 4e). In other words, methyl-plus TF motifs are more enriched in pCanyons, while methyl-minus TF motifs are more enriched in aCanyons. Therefore, when methylation increases, the aCanyons would lose TF binding, and the aCanyon targets would have decreased expression. On the other hand, when methylation increases, the pCanyons would have more TF binding, and the pCanyon targets would have increased expression.

Together, aCanyons and pCanyons, the two categories of methylation canyons defined by methylation competition, present distinct patterns of chromatin accessibility and gene regulation.

Discussion

Cytosine methylation is a reversible biochemical modification[35]. The global pattern of mammalian methylome is formed by two antagonizing processes: methylation and demethylation. For the first time, the methylation competition ratio reported in this paper utilizes methylation competing events inside the same DNA molecule to measure this antagonism. Through reanalyzing bisulfite sequencing data, we reveal that the methylation competition ratio is strongly correlated with gene expression. Different from the average methylation, the methylation competition ratio in different types of regulatory elements, e.g., promoter, gene-body, and enhancer, consistently has strong correlations with gene expression. Our analysis confirms the negative impact of methylation competition on gene transcription, which extends the interpretation of DNA methylation data. In terms of cancer gene regulation, we find that TSGs are characterized by conserved low methylation competition ratios across normal samples. During tumorigenesis, the elevation of methylation competition serves as a novel mechanism that can explain the repression of 40~60% of TSGs, which cannot be explained by hypermethylation. Thus, our definition of methylation competition expands our understanding of tumor methylation aberration in addition to promoter hypermethylation.

Polycomb repressive complex 2 (PRC2) is responsible for methylating histone H3 on Lys27 (H3K27)[36], a crucial chromatin mark for gene silencing in early development and oncogenesis[37]. Both single-molecule[38] and genome-wide[39] analysis reveal that PRC2 depositing is not favored by DNA methylation. Coincide with these findings,

the PRC2 catalytic subunit EZH2 (enhancer of Zeste homologue 2)[40], the stimulating subunit SUZ12 (suppressor of Zeste 12)[41], and H3K27 trimethylation (H3K27me3) are all negatively associated with the average methylation in promoters in human stem cells (Supplementary Fig 14b). However, they are all positive indicators of methylation competition (Supplementary Fig 14a). Bisulfite sequencing data in *Ezh2* conditional knockout mice[42] also confirm that the removal of PRC2 would cause more regions to have higher average methylation but lower methylation competition (Supplementary Fig 14c), suggesting that methylation competition can be promoted by PRC2 binding. Previous studies suggest that TET1 is associated with the repression of Polycomb targets[43,44], but they have not detected any direct interaction between PRC2 and TET1. Our analysis indicates that methylation competition may serve as the missing link in the PRC2-TET1 association.

Nucleosome positioning is essential for gene regulation by altering chromatin accessibility[45]. Nucleosome fuzziness measures the randomness of a nucleosome's position. Through reanalyzing human brain MNase-seq (micrococcal nuclease digestion with deep sequencing) data[46], we reveal that the methylation competition ratio is not associated with nucleosome fuzziness (Supplementary Fig 15), suggesting that methylation competition regulates gene expression independent of nucleosome positioning.

Methylation canyons are poorly methylated and have negligible differences in the average methylation, so they serve as an ideal context to investigate the methylation

competition. A previous study of mouse hematopoietic stem cells (HSCs) finds that some canyons are active while the others are silent, and that such a difference is explainable by the binding of H3K4me3 or H3K27me3[29]. Thanks to our definition of methylation competition, the distinct activities of different canyons can be explained by methylation data alone, suggesting that methylation competition indicates chromatin accessibility. Besides, we find that the active canyon (aCanyon) target genes present a unique regulating model: their negative correlation between gene expression and gene-body methylation (Fig 4d) has not been reported before. Although we have confirmed this observation by analyzing *in vitro* SELEX data of the 5mC preferences of TFs, more efforts are needed to scrutinize the mechanisms.

Recently, the competition between DNMTs and TETs is also confirmed in human ESCs[47]. With its different characteristics from the average methylation and methylation variation, the methylation competition ratio can disclose more methylation abnormality and cancer drivers. It highlights local methylation dynamics in the epigenetic landscape and will serve as a new layer of methylation biology.

Methods

Data source

All the public or controlled data used by this study are summarized in Supplementary Table 1. Totally 75 methylomes (WGBS and reduced representation bisulfite sequencing (RRBS)) were collected from the Roadmap Epigenomics project (http://www.roadmapepigenomics.org/), Encyclopedia of DNA Elements (ENCODE) project (https://www.encodeproject.org/), The Cancer Genome Atlas (TCGA) project (https://portal.gdc.cancer.gov/), Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) and DNA Data Bank of Japan (DDBJ, https://www.ddbj.nig.ac.jp) databases, including 69 human and 6 mouse datasets. Among them, there are eight matched normal-tumor pairs from TCGA, which are indicated by the TCGA patient ID. 44 RNA sequencing (RNA-seq) data were downloaded from Roadmap and TCGA, which are from the matched samples with methylomes. 38 ChIP-seq, 1 DNase-seq, 1 MNase-seq, and 1 Hi-C datasets were fetched from Roadmap or GEO.

Quality control and reads alignments

FastQC v0.11.7 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used for general quality checks of sequencing reads in FASTQ files. Then Trim Galore v0.6.4 (https://github.com/FelixKrueger/TrimGalore) was used to trim the sequencing adaptor and remove low-quality bases. WGBS reads were aligned to human (hg19) or mouse (mm9) genome using BSMAP v2.90[48] with default parameters. RRBS reads were also aligned by BSMAP, while an extra option '-D C-

CGG' was added to activate the RRBS mode[49]. The overlapping bases of two read mates were only counted once to avoid duplicate counting. RNA-seq reads were first mapped to human (hg19) genome by STAR v2.6.0c [50] with the option '--quantMode TranscriptomeSAM', and then gene expression level in FPKM (fragments per kilobase per million reads) was quantified using RSEM v1.3.1[51]. MNase-seq reads from the human brain were mapped to the human genome (hg19) by bowtie2 v2.2.7[52], then nucleosome binding positions and nucleosome fuzziness scores were called by DANPOS v2.2.2[53] with the 'dpos' command. The Hi-C reads from CD3+ T-cell were mapped to the human genome (hg19) by bowtie2. Then the chromatin interactions under 2-kilobase resolution were called using the 'analyzeHiC' module in Homer v4.8[54].

Genome annotation and genomic features

The NCBI RefSeq gene annotation files (hg19 and mm9) were fetched using the Table Browser tool in the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTables). Gene promoter regions were defined as 1kb upstream of TSS to 500bp downstream of TSS. Promoters' CpG ratios were calculated by this formula: (number of CpGs × number of bp) / (number of Cs × number of Gs) [23]. Then promoters were stratified into three groups according to their CpG ratios, i.e., high-CpG promoters (HCPs), intermedium-CpG promoters (ICPs), and low-CpG promoters (LCPs). Gene-body regions were defined as 500bp downstream of TSS to the transcription termination site (TTS). CpG-island positions were fetched using the UCSC Table Browser, which was derived based on the published formula[55].

Enhancers and their target genes in CD3+ T-cell were defined based on Hi-C data. First, the pairs of anchors which have chromatin interaction were annotated to the genome. Then for each pair, if only one of them located in promoter or gene-body regions, it was assigned as an enhancer-target pair.

The quantification of competition between active DNA methylation and demethylation

The methylation competition events are represented by the unmethylated CpGs in partially methylated reads (slashed circles in Fig 1a). So, bisulfite sequencing reads were dissected into three categories of fragments, i.e., methylated fragments (consecutive solid circles in Fig 1a), unmethylated fragments (consecutive blank circles in Fig 1a), and methylation competition fragments (consecutive slashed circles in Fig 1a). Thus, the methylation competition ratio of a particular genomic region was measured by the following equation.

$$\text{methylation competition ratio} = \frac{\sum_{c=1}^{C} \omega_c}{\sum_{c=1}^{C} \omega_c + \sum_{m=1}^{M} \omega_m + \sum_{u=1}^{U} \omega_u} \qquad \text{(Eq. 1)}$$

M, U, and C represent the numbers of methylated, unmethylated, and methylation-competition fragments, respectively. ω_m, ω_u, and ω_c are the weights for each fragment. They can be set as the CpG counts of each fragment (weighted) or 1 (unweighted). Comparing the weighted and unweighted ratios, we found that the weighted ratio (Fig 2b; left panel) has slightly better correlations with gene expression than the unweighted ratio does (Supplementary Fig 4). Hence, the weighted methylation competition ratio was used in this study. Note that the weighted methylation competition ratio is equivalent to the proportion of CpGs in

methylation-competition fragments among all the CpGs. Hence, the methylation competition ratio is also applicable to single CpG site. Theoretically, methylation competition ratios vary between 0 and 1 (not equal to 1). But in real data, as more than half of reads are either fully methylated or fully unmethylated (Supplementary Fig 16a), methylation competition ratios of most regions are lower than 0.5 (Supplementary Fig 16b and 16c and 16d). To reduce false positive, we only retained CpGs covered by at least four reads in the analysis. We integrated the calculation of the methylation competition ratio and the average methylation ratio in a single Python script available at https://github.com/JiejunShi/methylation_competition.

Quantifications of the average methylation

Two measures of the average methylation are used in this study: traditional mean methylation and cellular heterogeneity-adjusted mean methylation (CHALM)[21]. The traditional mean methylation of a region is calculated as the proportion of methylated CpGs among all CpGs covered by reads. The CHALM value of a region is computed as $n_m/(n_m + n_u)$, where n_m and n_u are the counts of methylated reads and unmethylated reads, respectively.

Visualization of the methylation competition ratio and the average methylation

Methylation competition ratios and average methylation measures at base-pair resolution were generated in the 'wiggle' format using our 'methylation competition' tool (https://github.com/JiejunShi/methylation_competition). The 'wigToBigWig' program from UCSC binary utility directory

(http://hgdownload.soe.ucsc.edu/admin/exe/) was used to transform 'wiggle' files to 'bigWig' files. For viewing specific gene loci, the 'bigWig' files were uploaded to 'custom tracks' in UCSC Genome Browser (http://genome.ucsc.edu/), with genome assembly of hg19 or mm9. The 'computeMatrix' module of deepTools v3.2.1[56] was used to extract the scores of interested regions into matrix format. Then the averaged profiles and heatmaps were visualized using 'plot' and 'image' functions in R.

The co-localization analysis of DNMT3A1 and TET1

The ChIP-seq reads density profiles of DNMT3A1 and TET1 in mouse ESCs were downloaded from the GEO database (GSE100951 and GSE100955). They have been corrected for input background using DANPOS. Then the reads densities in promoter regions were extracted and transformed into log2 RPKM values by normalizing for promoter lengths and total reads counts. Hence, the promoter binding intensities of DNMT3A1 or TET1 were defined as the standardized values $\pi = \frac{x-\mu}{\sigma}$, where π is the binding intensity, x is the log2 RPKM, μ is the mean of x across promoters, and σ is the standard deviation of x across promoters. Then the 'DNMT3A1-TET1 joint regulation score' (Π) of a promoter region is defined in the following equation.

$$\Pi = \frac{\pi^D - \pi^D_{min}}{\pi^D_{max} - \pi^D_{min}} \cdot \frac{\pi^T - \pi^T_{min}}{\pi^T_{max} - \pi^T_{min}}, \qquad \text{(Eq. 2)}$$

where π^D and π^T represent the respective binding intensities of DNMT3A1 and TET1 in the promoter region; π^D_{min} and π^T_{min} indicate the minimum of π^D and π^T

across promoters; π^D_{max} and π^T_{max} indicate the maximum of π^D and π^T across promoters. By definition, Π is between 0 and 1.

Quantifications of the methylation variation

DNA methylation variation is quantified by the 'methylation heterogeneity'(MH) scores. In the MH algorithms, reads cover fewer than 4 CpGs are excluded. Same as the quantification of the methylation competition, only CpGs covered by at least four reads were included in the analysis to reduce false positive. The proportion of discordant reads (PDR) was calculated as previously described[12]. PDR for a 4-CpG locus was the proportion of partially methylated reads among all reads which cover the four adjacent CpGs. Epipolymorphism[13] for a 4-CpG locus was calculated as $1 - \sum(p_i^2)$, where p_i is the frequency of all possible methylation patterns of this locus. Methylation entropy for a 4-CpG locus is measured by the equation of Shannon's entropy $-\sum(p_i \ln p_i)$, where p_i is the same as in Epipolymorphism. The entropy/ Epipolymorphism/PDR of a region is calculated as the average value of all 4-CpG locus in that region. The entropy value of 4-CpG loci ranges from 0 to 4. So, a cutoff of 0.4 is used to define the notable entropy change. A comparison study[57] has implemented the above MH algorithms, and the codes are available on GitHub (https://github.com/MPIIComputationalEpigenetics/WSHScripts).

Differentially methylated regions (DMRs) and differential methylation-competition regions (DMCRs)

DMRs and DMCRs are identified by Metilene v0.2-8[58] in 'de novo' mode with the option of '-m 5' to get regions of at least 5 CpGs. For DMRs (hypermethylated and

hypomethylated regions), '-d 0.4' is used to detect the regions with at least 40% of average methylation difference. For DMCRs (competition-elevated and competition-depleted regions), because methylation competition scores of most regions are lower than 0.5 (Supplementary Fig 16), '-d 0.2' is applied to find regions with at least 20% of methylation competition difference. The p-value cutoff for both DMR and DMCR is 0.05.

Gene set enrichment analysis (GSEA) and gene ontology (GO) analysis

GSEA was applied by GSEA software v4.0.3[59] in the 'pre-ranked' mode with default parameters. The genes were decreasingly sorted by the promoter methylation competition ratio. All the functional terms are collected in GSEA Molecular Signatures Database (MsigDB, v7.1), except 'COSMIC Tumor Suppressors', 'COSMIC Oncogenes', and 'PANCAN Driver Genes'. 'COSMIC Tumor Suppressors' and 'COSMIC Oncogenes' were fetched from COSMIC Cancer Gene Census (CGC). 'PANCAN Driver Genes' were downloaded from a previous study[60], which integrated cancer drivers from 33 cancer types. The normalized enrichment scores (NES) and FDR values were calculated by GSEA software. The heatmap of NES was plotted using the 'image' function in R. And the running enrichment scores were re-plotted using codes published on GitHub (https://github.com/PeeperLab/Rtoolbox/blob/master/R/ReplotGSEA.R). The GO analysis of methylation altered genes was conducted using the DAVID online tool (https://david.ncifcrf.gov/)[61]. And the Benjamini-Hochberg adjusted p-values were calculated by DAVID. The functional enrichment of methylation canyon target genes

was measured by two-tailed Fisher's Exact Test, and the p-values were adjusted using the Benjamini-Hochberg method.

Defining undermethylated regions (UMRs)

UMRs were detected using the published method[29] based on the Hidden Markov model. To reduce false positive, we only retained CpGs covered by at least four reads for UMR detection. The CpGs with an average methylation ratio lower than 10% were defined as undermethylated CpGs. Then UMRs were identified as regions that include at least four consecutive undermethylated CpGs. The adjacent UMRs were merged into a single UMR if the average methylation ratio of the merged UMR was still lower than 10%.

Determining the threshold for categorizing methylation canyons

Methylation canyons are the large (≥3.5kb) and conserved UMRs[29,31,32]. So, the UMRs longer than 3.5kb were identified as canyons. And the canyon target genes were defined as genes whose promoter or gene-body region is overlapped with canyons. To set the threshold for canyon categorizing, we calculated methylation competition ratios of 10,000 non-overlapped random genomic regions to determine the distribution of genome backgrounds. Then 'fitdist' function in R package 'fitdistrplus' was used to fit it to the classical distributions, including 'normal', 'log-normal', 'beta', 'gamma', 'uniform', 'exponential', and 'logistic' distributions (Supplementary Fig 9a). The parameters for each distribution were estimated by 'maximum likelihood estimation'. By using the Cramér–von Mises criterion, we found the fitness to 'gamma' has the minimum distance with the distribution of genome

background methylation competition ratios, which means ‘gamma’ distribution is the best fit (Supplementary Fig 9b). And this was also confirmed by the Q-Q plot (Supplementary Fig 9c). Given the estimated parameters ‘shape’ and ‘rate’ from ‘gamma’ fitness, the threshold was determined as the 90 percentiles of the background distribution, which is aiming to select the regions with a significantly higher methylation competition ratio. Then canyons were categorized based on this threshold. Canyons with methylation competition ratio higher than this threshold were assigned as ‘pCanyons’ (Polycomb canyons), while the rest were designated as ‘aCanyons’ (active canyons) (Supplementary Fig 9d).

Locus-specific correlation between average methylation and gene expression

The gene promoter and downstream regions (2kb upstream to 10kb downstream of TSS) were equally divided into 120 bins. The average methylation ratio was calculated for each bin, which is in 100 bp length. Aligning the TSS of different genes, all the average methylation ratios were organized into a matrix, in which rows are genes and columns are bins. Then the Spearman’s rank correlation coefficients were computed between the gene expression vector and each column of this matrix. The resulting vectors of correlation coefficients were visualized as bar plots.

Enrichment of motifs with different 5mC preference

The 5mC preferences of TF motifs were quantified by the SELEX method introduced by the previous study[34]. The motifs with preference values (termed as ‘mCG enrichment’ in [34]) higher than 0 were assigned as ‘methyl-plus’ motifs, whose binding can be enhanced by 5mC. While motifs with preference value lower than 0 were

defined as ‘methyl-minus’ motifs, whose binding is not favored by 5mC. Besides these two categories, there are also motifs with multiple effects or little effect on 5mC. Removing these ambiguous motifs and motifs with identical sequences, we got 105 non-redundant motifs that are clearly ‘methyl-minus’ or ‘methyl-plus’. Then the ‘findMotifsGenome’ module in Homer software was used to call the motif positions in canyon sequences. The enrichment score of a particular motif was calculated as motif counts per kilobase (CPK) of the canyon. Then the fold enrichment of this motif was defined as the fold change between its CPK at aCanyons and CPK at pCanyons. Spearman’s rank correlation between fold enrichment and 5mC preference was visualized in a scatter plot. And the p-value was calculated by the correlation test.

Transcription factor binding difference between methylation canyon groups

Positions of TF binding peaks determined in H1 human stem cells were downloaded from the Roadmap website. Then the percentages of TF-occupied aCanyons and pCanyons were calculated based on whether the canyon overlaps with TF binding peaks. The binding difference was measured by the odds ratio between these two percentages, i.e., TF occupied aCanyons (%) / TF occupied pCanyons (%). TFs with odds ratios higher than 1 are more enriched in aCanyons, while the others are more enriched in pCanyons.

Chromatin interaction at methylation canyons

Chromatin interactions were defined by pairs of Hi-C anchors. Then the overlap between anchor pairs and canyons was detected using bedtools v2.25.0[62]. As a

control, genomic regions have the same length as canyons, but random genomic positions were generated using bedtools. For each anchor pairs, if both of them are located in the same canyon or random genomic region, it is called ‘self-interacting’. If only one of them is located in a canyon or random region, it is called ‘distant-interacting’.

Code availability

The open-source software for methylation competition is freely available at https://github.com/JiejunShi/methylation_competition.

Acknowledgements

We thank members of the Li lab for helpful discussions. This work was supported by the US National Institutes of Health to W.L. (R01HG007538, R01CA193466, R01CA228140).

Author contributions

J.S. and W.L. conceived and developed the outline of this research. J.S. wrote the tools and performed data analysis and method evaluations. J.X., Y.C., J.L., and L.S. assisted with the manuscript. J.S. and W.L. wrote the paper.

Competing interests

After completing the current studies at Baylor College of Medicine, J.X. became a full-time employee at Helio Health. W.L. is a consultant for Helio Health and ChosenMed. The remaining authors declare no competing interests.

References

1. Hermann, A., Goyal, R. & Jeltsch, A. The Dnmt1 DNA-(cytosine-C5)-methyltransferase Methylates DNA Processively with High Preference for Hemimethylated Target Sites. *J Biol Chem* **279**, 48350–48359 (2004).

2. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. *Cell* **99**, 247–257 (1999).

3. Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**, 1129–33 (2010).

4. Tahiliani, M. *et al.* Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Sci New York N Y* **324**, 930–5 (2009).

5. Verma, N. *et al.* TET proteins safeguard bivalent promoters from de novo methylation in human embryonic stem cells. *Nat Genet* **50**, 83–95 (2017).

6. Gu, T. *et al.* DNMT3A and TET1 cooperate to regulate promoter epigenetic landscapes in mouse embryonic stem cells. *Genome Biol* **19**, 88 (2018).

7. Zhang, X. *et al.* DNMT3A and TET2 compete and cooperate to repress lineage-specific transcription factors in hematopoietic stem cells. *Nat Genet* **48**, 1014 1023 (2016).

8. Couronné, L., Bastard, C. & Bernard, O. A. TET2 and DNMT3A mutations in human T-cell lymphoma. *New Engl J Medicine* **366**, 95–6 (2012).

9. Herman, J. G. & Baylin, S. B. Gene Silencing in Cancer in Association with Promoter Hypermethylation. *New Engl J Medicine* **349**, 2042–2054 (2003).

10. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–22 (2009).

11. Hon, G. C. *et al.* Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nat Genet* **45**, 1198 1206 (2013).

12. Landau, D. A. *et al.* Locally Disordered Methylation Forms the Basis of Intratumor Methylome Variation in Chronic Lymphocytic Leukemia. *Cancer Cell* **26**, 813–825 (2014).

13. Landan, G. *et al.* Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat Genet* **44**, 1207 (2012).

14. Guo, S. *et al.* Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat Genet* **49**, 635–642 (2017).

15. Jenkinson, G., Pujadas, E., Goutsias, J. & Feinberg, A. P. Potential energy landscapes identify the information-theoretic nature of the epigenome. *Nat Genet* **49**, 719–729 (2017).

16. Gaiti, F. *et al.* Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature* **569**, 576–580 (2019).

17. Hernando-Herraez, I. *et al.* Ageing affects DNA methylation drift and transcriptional cell-to-cell variability in mouse muscle stem cells. *Nat Commun* **10**, 4361 (2019).

18. Ginno, P. A. *et al.* A genome-scale map of DNA methylation turnover identifies site-specific dependencies of DNMT and TET activity. *Nat Commun* **11**, 2680 (2020).

19. Zhu, W.-G. *et al.* Methylation of Adjacent CpG Sites Affects Sp1/Sp3 Binding and Activity in the p21Cip1 Promoter. *Mol Cell Biol* **23**, 4056–4065 (2003).

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

21. Xu, J. *et al.* Cellular heterogeneity–adjusted clonal methylation (CHALM) provides better prediction of gene expression. *Biorxiv* 2020.02.23.961813 (2020) doi:10.1101/2020.02.23.961813.

22. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317 330 (2015).

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

24. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18**, 696–705 (2018).

25. Scimone, M. L., Cote, L. E. & Reddien, P. W. Orthogonal muscle fibres have different instructive roles in planarian regeneration. *Nature* **551**, 623–628 (2017).

26. Esteller, M. CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene* **21**, 5427–5440 (2002).

27. Cao, J. *et al.* PLZF mediates the PTEN/AKT/FOXO3a signaling in suppression of prostate tumorigenesis. *Plos One* **8**, e77922 (2013).

28. Robinson, D. *et al.* Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215–28 (2015).

29. Jeong, M. *et al.* Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nat Genet* **46**, 17 23 (2014).

30. Xie, W. *et al.* Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134–48 (2013).

31. Zhang, X. *et al.* Large DNA Methylation Nadirs Anchor Chromatin Loops Maintaining Hematopoietic Stem Cell Identity. *Mol Cell* **78**, 506-521.e6 (2020).

32. Su, J. *et al.* Homeobox oncogene activation by pan-cancer DNA hypermethylation. *Genome Biol* **19**, 108 (2018).

33. Ball, M. P. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat Biotechnol* **27**, 361–8 (2009).

34. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017).

35. Ramchandani, S., Bhattacharya, S. K., Cervoni, N. & Szyf, M. DNA methylation is a reversible biological signal. *Proc National Acad Sci* **96**, 6107–6112 (1999).

36. Simon, J. A. & Kingston, R. E. Mechanisms of polycomb gene silencing: knowns and unknowns. *Nat Rev Mol Cell Bio* **10**, 697 708 (2009).

37. Sparmann, A. & Lohuizen, M. van. Polycomb silencers control cell fate, development and cancer. *Nat Rev Cancer* **6**, 846–856 (2006).

38. Murphy, P. J. *et al.* Single-molecule analysis of combinatorial epigenomic states in normal and tumor cells. *P Natl Acad Sci Usa* **110**, 7772–7 (2013).

39. Li, Y. *et al.* Genome-wide analyses reveal a role of Polycomb in promoting hypomethylation of DNA methylation valleys. *Genome Biol* **19**, 18 (2018).

40. Margueron, R. *et al.* Ezh1 and Ezh2 maintain repressive chromatin through different mechanisms. *Mol Cell* **32**, 503–18 (2008).

41. Cao, R. & Zhang, Y. SUZ12 Is Required for Both the Histone Methyltransferase Activity and the Silencing Function of the EED-EZH2 Complex. *Mol Cell* **15**, 57–67 (2004).

42. Sashida, G. *et al.* Ezh2 loss promotes development of myelodysplastic syndrome but attenuates its predisposition to leukaemic transformation. *Nat Commun* **5**, 4177 (2014).

43. Wu, H. *et al.* Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature* **473**, 389–393 (2011).

44. Williams, K. *et al.* TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343–348 (2011).

45. Jiang, C. & Pugh, B. F. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet* **10**, 161 172 (2009).

46. Li, Y. *et al.* Human exonization through differential nucleosome occupancy. *P Natl Acad Sci Usa* **115**, 8817–8822 (2018).

47. Charlton, J. *et al.* TETs compete with DNMT3 activity in pluripotent cells at thousands of methylated somatic enhancers. *Nat Genet* 1–9 (2020) doi:10.1038/s41588-020-0639-9.

48. Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *Bmc Bioinformatics* **10**, 232 (2009).

49. Xi, Y. *et al.* RRBSMAP: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing. *Bioinform Oxf Engl* **28**, 430–2 (2011).

50. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

51. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *Bmc Bioinformatics* **12**, 323 (2011).

52. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–9 (2012).

53. Chen, K. *et al.* DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res* **23**, 341–51 (2012).

54. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576–89 (2010).

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

56. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-5 (2016).

57. Scherer, M. *et al.* Quantitative comparison of within-sample heterogeneity scores for DNA methylation data. *Nucleic Acids Res* (2020) doi:10.1093/nar/gkaa120.

58. Jühling, F. *et al.* metilene: fast and sensitive calling of differentially methylated regions from bisulfite sequencing data. *Genome Res* **26**, 256–62 (2015).

59. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc National Acad Sci* **102**, 15545–15550 (2005).

60. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385.e18 (2018).

61. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44–57 (2009).

62. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinform Oxf Engl* **26**, 841–2 (2010).

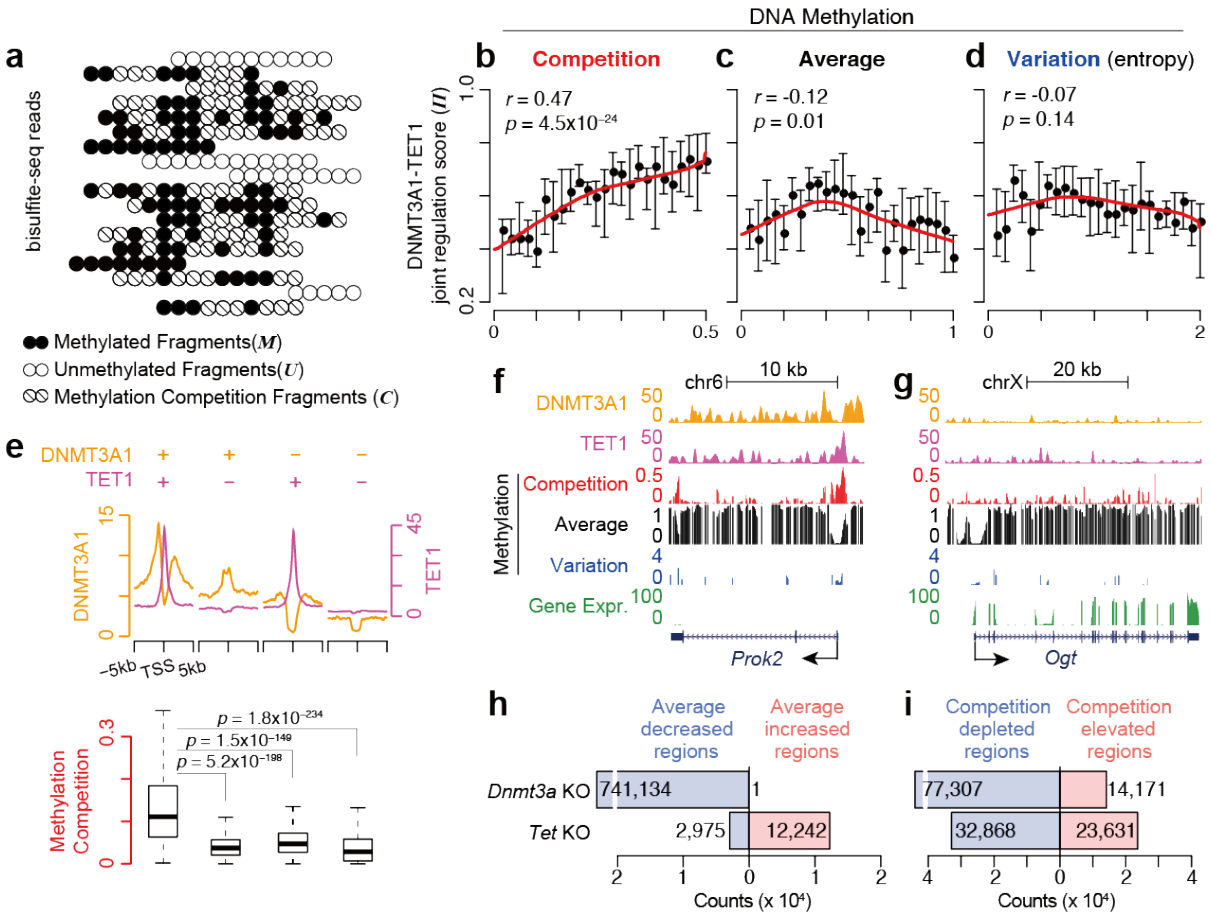

Figure 1. Methylation competition ratio measures the antagonism between methylation and demethylation processes in mESCs

(**a**) Schematic of active DNA methylation competition captured by bisulfite-seq. Solid circles are methylated cytosines. Blank circles are unmethylated cytosines. Slashed circles are unmethylated cytosines in partially methylated reads, i.e., methylation-competition cytosines. (**b**) Methylation competition is positively correlated with the 'DNMT3A1-TET1 joint regulation score' (Π) in gene promoter regions. Average methylation (**c**) and methylation variation (**d**) are not correlated with Π at gene promoters. Spearman's rank correlation was calculated based on all data points. P-values were calculated by the two-tailed correlation test for Spearman's correlation. For visualization, gene promoters were divided into 25 groups based on the methylation scores in (b), (c), and (d). LOWESS lines were plotted to describe the relationships between variables (indicated by red curves). (**e**) The methylation competition ratio is significantly higher at DNMT3A1&TET1 co-occupied promoters. Gene numbers of each groups: 'DNMT3A1+TET1+', n=1,294; 'DNMT3A1+TET1-', n=1,238; 'DNMT3A1-TET1+', n=1,370; 'DNMT3A1-TET1-', n=1,595. The two-tailed Mann-Whitney U test was used for the significance test. (**f**) UCSC Genome Browser tracks show DNMT3A1 binding (orange), TET1 binding (purple), methylation competition (red), average methylation (black), methylation variation (blue), and gene expression data (green) at *Prok2* gene. (**g**) Same as (f), but for gene *Ogt*. (**h**) *Dnmt3a* knockout leads to decrease of average methylation, while *Tet* knockout leads to hypermethylation. (**i**) Both *Dnmt3a* knockout and *Tet* knockout lead to more

competition depletion than elevation. '*Dnmt3a* KO' means *Dnmt3a1* and *Dnmt3a2* double knockout mESC. '*Tet* KO' means *Tet1, Tet2,* and *Tet3* triple knockout mESC.

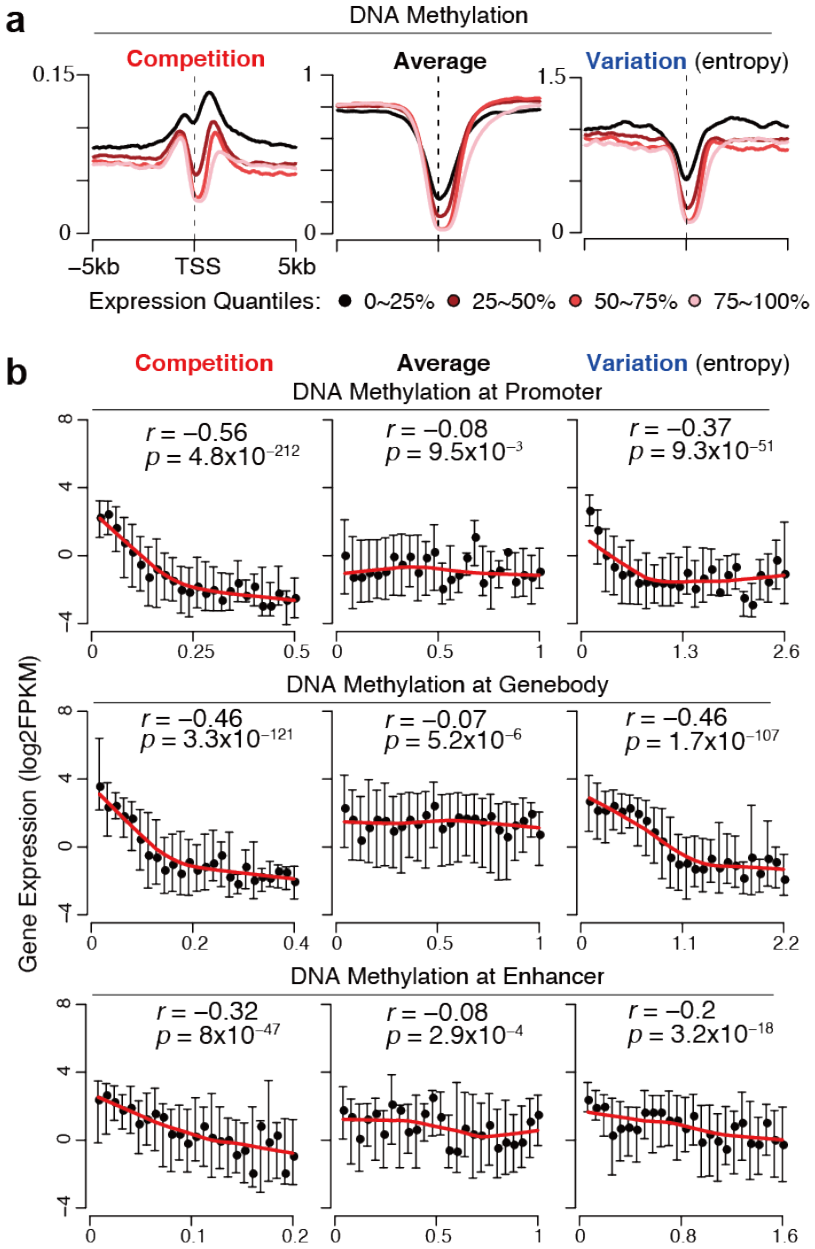

Figure 2. Methylation competition is negatively associated with gene expression

(**a**) Average profiles in TSS-proximal regions of 4 gene groups, which are divided by quantiles of expression levels, in CD3+ T-cells. Methylation competition ratios are on the left, average methylation ratios are on the middle, and methylation variation scores(entropy) are on the right. 0~25%, lowest expressed; 25~50%, lower expressed; 50~75%, higher expressed; 75~100%, highest expressed. The gene number of each group is 3,600. (**b**) The promoter/gene-body/enhancer methylation competition ratios (1st column) are strongly negatively correlated with gene expression level in CD3+ T-cells, and this correlation is stronger than that of the average methylation (2nd column) and the methylation variation (3rd column). Promoter regions are from 1kb upstream to 500bp downstream of TSS. Gene-body regions are from 500bp downstream of TSS to TTS. Enhancer regions are defined

based on chromatin interactions validated by Hi-C data (see Methods). Spearman's rank correlation was calculated based on all data points. P-values were calculated by the two-tailed correlation test for Spearman's correlation. For visualization, genomic regions (e.g., promoter, gene-body, and enhancer) were divided into 25 groups based on each quantitative measure of methylation. LOWESS lines were plotted to describe the relationships between variables (indicated by red curves).

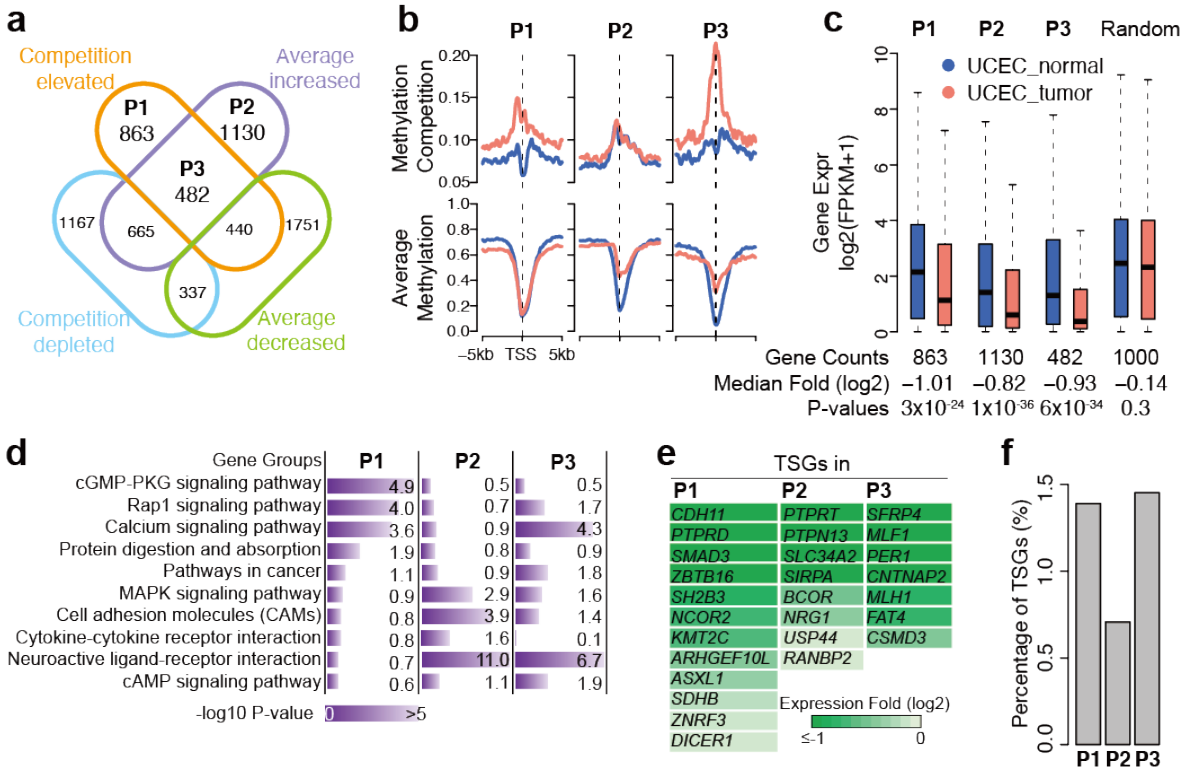

Figure 3. Tumor repressors can be repressed by methylation competition elevation rather than hypermethylation in uterus tumor

(**a**) Venn diagram shows the overlap between promoter methylation competition altered genes and average methylation altered genes in TCGA uterus tumor (UCEC). P1, methylation competition-elevated but average methylation stable genes; P2, hypermethylated but competition-stable genes; P3, competition-elevated and hypermethylated genes. (**b**) Average profiles of methylation competition (upper panel) and average methylation (lower panel) in TSS regions of P1, P2, and P3 genes. UCEC normal is in blue, and UCEC tumor is in red. (**c**) Gene expression change of P1, P2, P3, and 1,000 randomly selected genes. The fold changes between median values (log2 scale) are indicated below. The two-tailed Wilcoxon signed-rank test was used for the significance test. The lists of repressed genes in 3 groups are in Supplementary Table 2. (**d**) Gene ontology analysis on P1, P2, and P3 genes using DAVID online tool. The Benjamini-Hochberg adjusted p-values were given by DAVID. (**e**) COSMIC tumor suppressor genes which are overlapped with P1, P2, and P3 genes. Gene expression fold changes (log2 scale) are indicated. Darker green means higher repression in the UCEC tumor sample. (**f**) The percentages of tumor suppressors in P1, P2, and P3 genes.

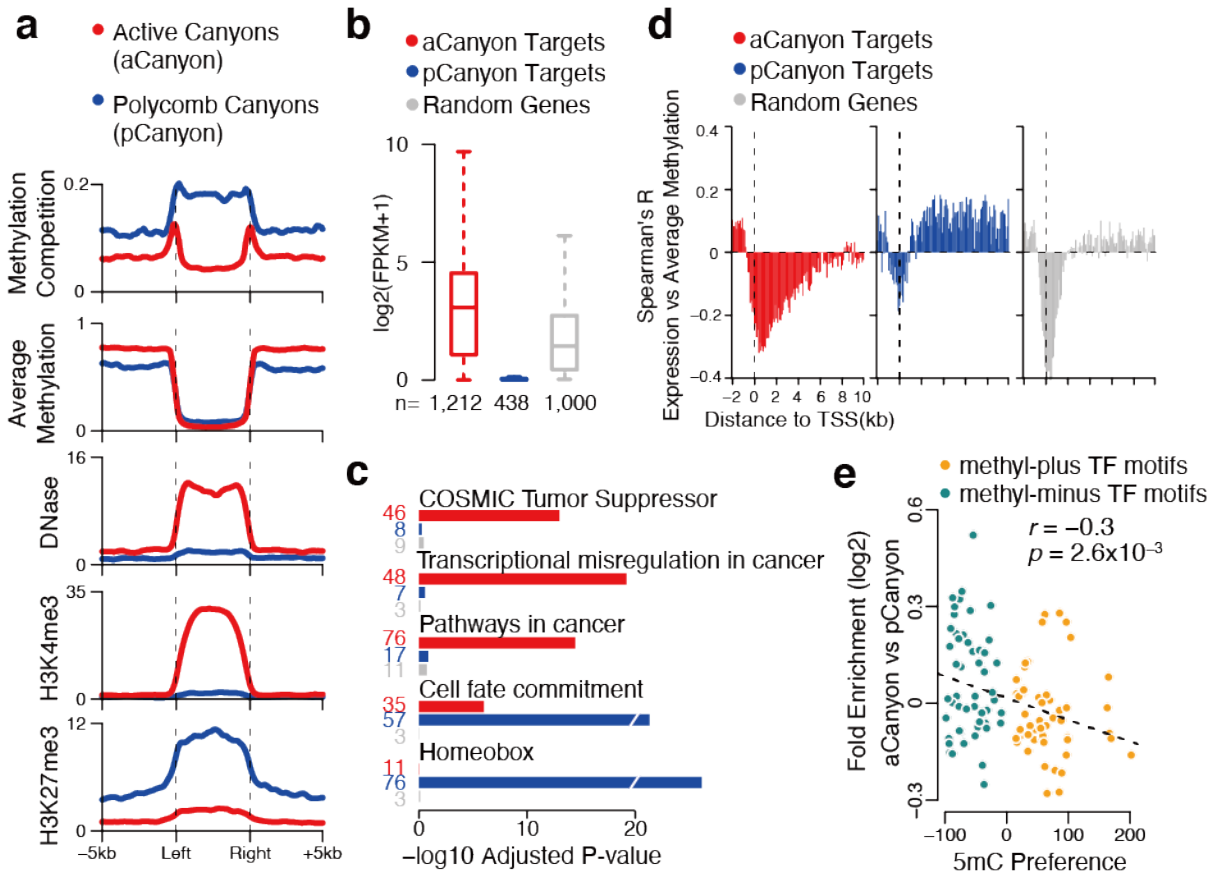

Figure 4. Distinct characteristics of methylation canyons categorized by methylation competition

(**a**) Average profiles of methylation competition, average methylation, DNase I hypersensitive sites, H3K4me3, and H3K27me3 in methylation canyons in CD3+ T-cell. 'aCanyons' (red) are low-competition canyons. 'pCanyons' (blue) are high-competition canyons. The X-axis indicates the distance to canyon borders. And all canyons are normalized to 5kb length. (**b**) Expression of aCanyon target genes (red), pCanyon target genes (blue), and 1,000 randomly selected genes (grey). Lists of aCanyons and pCanyons are in Supplementary Table 3, together with their target genes. (**c**) Functional enrichment analysis of aCanyon target genes, pCanyon target genes, and 1,000 randomly selected genes. Enriched gene counts in each group are indicated on the left side. P-values were measured by two-tailed Fisher's Exact Test and adjusted by the Benjamini-Hochberg method. (**d**) Spearman correlation between gene expression and average methylation of 100bp-bin in gene regions. Totally 120 bins from -2kb to 10kb were measured. (**e**) Relationship between fold enrichment and 5mC preference of TF motifs. Each dot represents a motif. Y-axis indicates the fold change (log2) between enrichment at aCanyon and enrichment at pCanyon of the same motif (see Methods). The X-axis shows the 5mC preference of motifs measured by the SELEX technique. 'methyl-plus' TFs prefer to bind methylated sequences, while binding of 'methyl-minus' TFs are not favored by 5mC. The list of 'methyl-plus' and 'methyl-minus' TFs are in Supplementary Table 4. Spearman's rank correlation was used. P-values were calculated by the two-tailed correlation test

for Spearman’s correlation. The linear model was plotted to describe the relationships between variables (indicated by the dashed line).

Figures

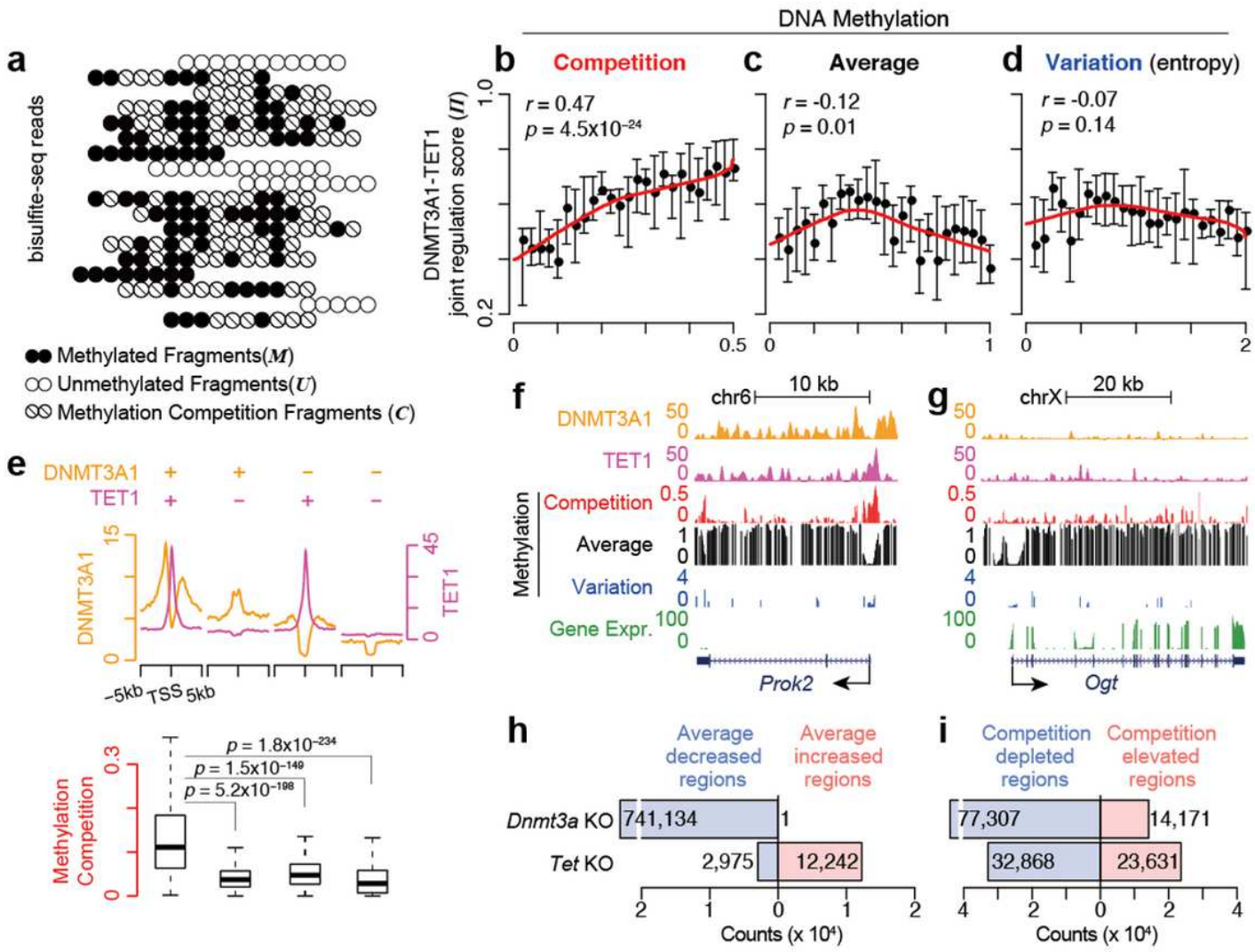

Figure 1

Methylation competition ratio measures the antagonism between methylation and demethylation processes in mESCs (a) Schematic of active DNA methylation competition captured by bisulfite-seq. Solid circles are methylated cytosines. Blank circles are unmethylated cytosines. Slashed circles are unmethylated cytosines in partially methylated reads, i.e., methylation- competition cytosines. (b) Methylation competition is positively correlated with the 'DNMT3A1-TET1 joint regulation score' (Π) in gene promoter regions. Average methylation (c) and methylation variation (d) are not correlated with Π at gene promoters. Spearman's rank correlation was calculated based on all data points. P-values were calculated by the two-tailed correlation test for Spearman's correlation. For visualization, gene promoters were divided into 25 groups based on the methylation scores in (b), (c), and (d). LOWESS lines were plotted to describe the relationships between variables (indicated by red curves). (e) The methylation competition ratio is significantly higher at DNMT3A1&TET1 co-occupied promoters. Gene numbers of each groups: 'DNMT3A1+TET1+', n=1,294; 'DNMT3A1+TET1-', n=1,238; 'DNMT3A1-TET1+', n=1,370;

'DNMT3A1-TET1-', n=1,595. The two-tailed Mann-Whitney U test was used for the significance test. (f) UCSC Genome Browser tracks show DNMT3A1 binding (orange), TET1 binding (purple), methylation competition (red), average methylation (black), methylation variation (blue), and gene expression data (green) at Prok2 gene. (g) Same as (f), but for gene Ogt. (h) Dnmt3a knockout leads to decrease of average methylation, while Tet knockout leads to hypermethylation. (i) Both Dnmt3a knockout and Tet knockout lead to more competition depletion than elevation. 'Dnmt3a KO' means Dnmt3a1 and Dnmt3a2 double knockout mESC. 'Tet KO' means Tet1, Tet2, and Tet3 triple knockout mESC.

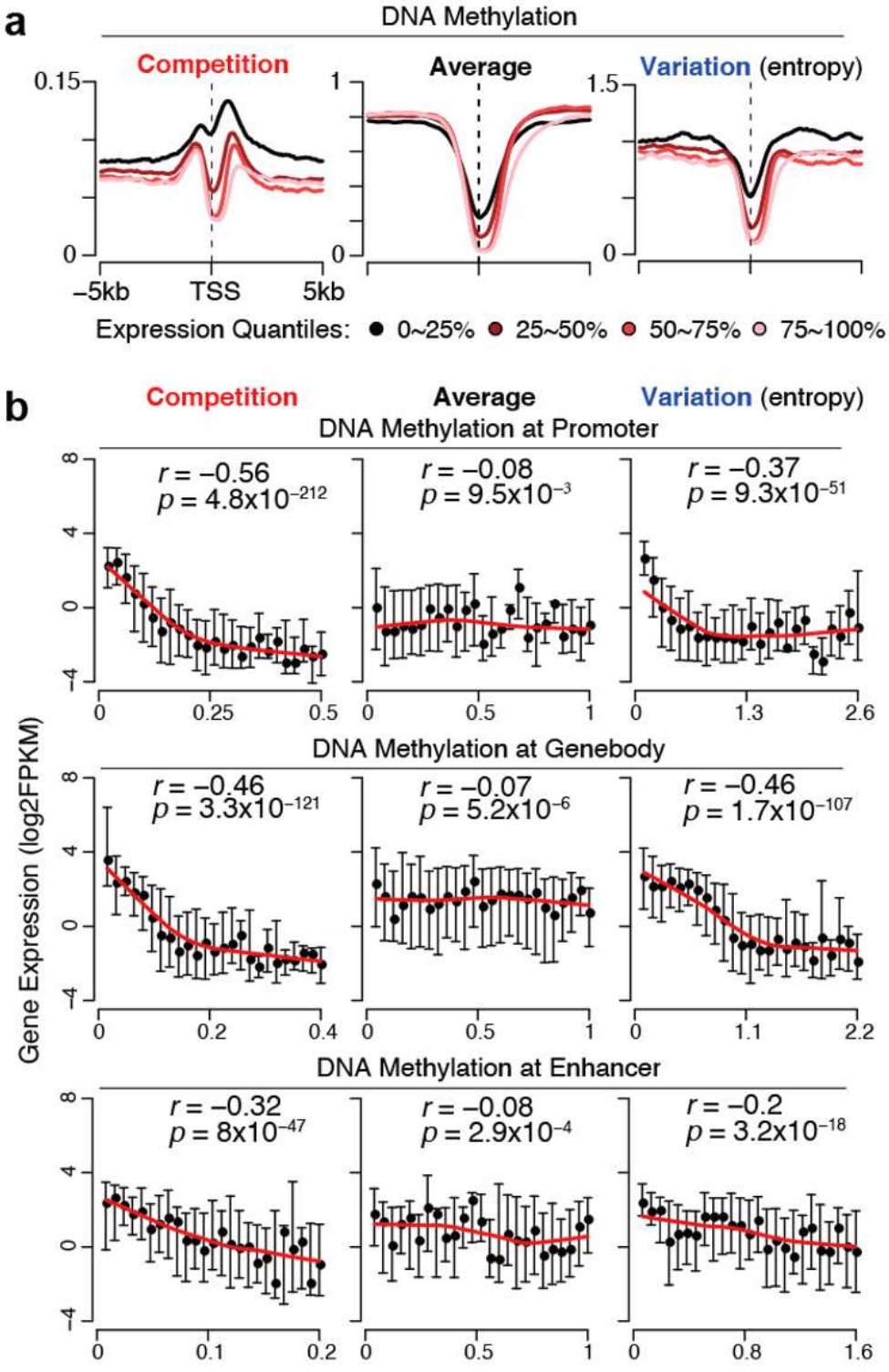

Figure 2

Methylation competition is negatively associated with gene expression (a) Average profiles in TSS-proximal regions of 4 gene groups, which are divided by quantiles of expression levels, in CD3+ T-cells. Methylation competition ratios are on the left, average methylation ratios are on the middle, and methylation variation scores(entropy) are on the right. 0~25%, lowest expressed; 25~50%, lower expressed; 50~75%, higher expressed; 75~100%, highest expressed. The gene number of each group is 3,600. (b) The promoter/gene-body/enhancer methylation competition ratios (1st column) are strongly negatively correlated with gene expression level in CD3+ T-cells, and this correlation is stronger than that of the average methylation (2nd column) and the methylation variation (3rd column). Promoter regions are from 1kb upstream to 500bp downstream of TSS. Gene-body regions are from 500bp downstream of TSS to TTS. Enhancer regions are defined based on chromatin interactions validated by Hi-C data (see Methods). Spearman's rank correlation was calculated based on all data points. P-values were calculated by the two-tailed correlation test for Spearman's correlation. For visualization, genomic regions (e.g., promoter, gene-body, and enhancer) were divided into 25 groups based on each quantitative measure of methylation. LOWESS lines were plotted to describe the relationships between variables (indicated by red curves).

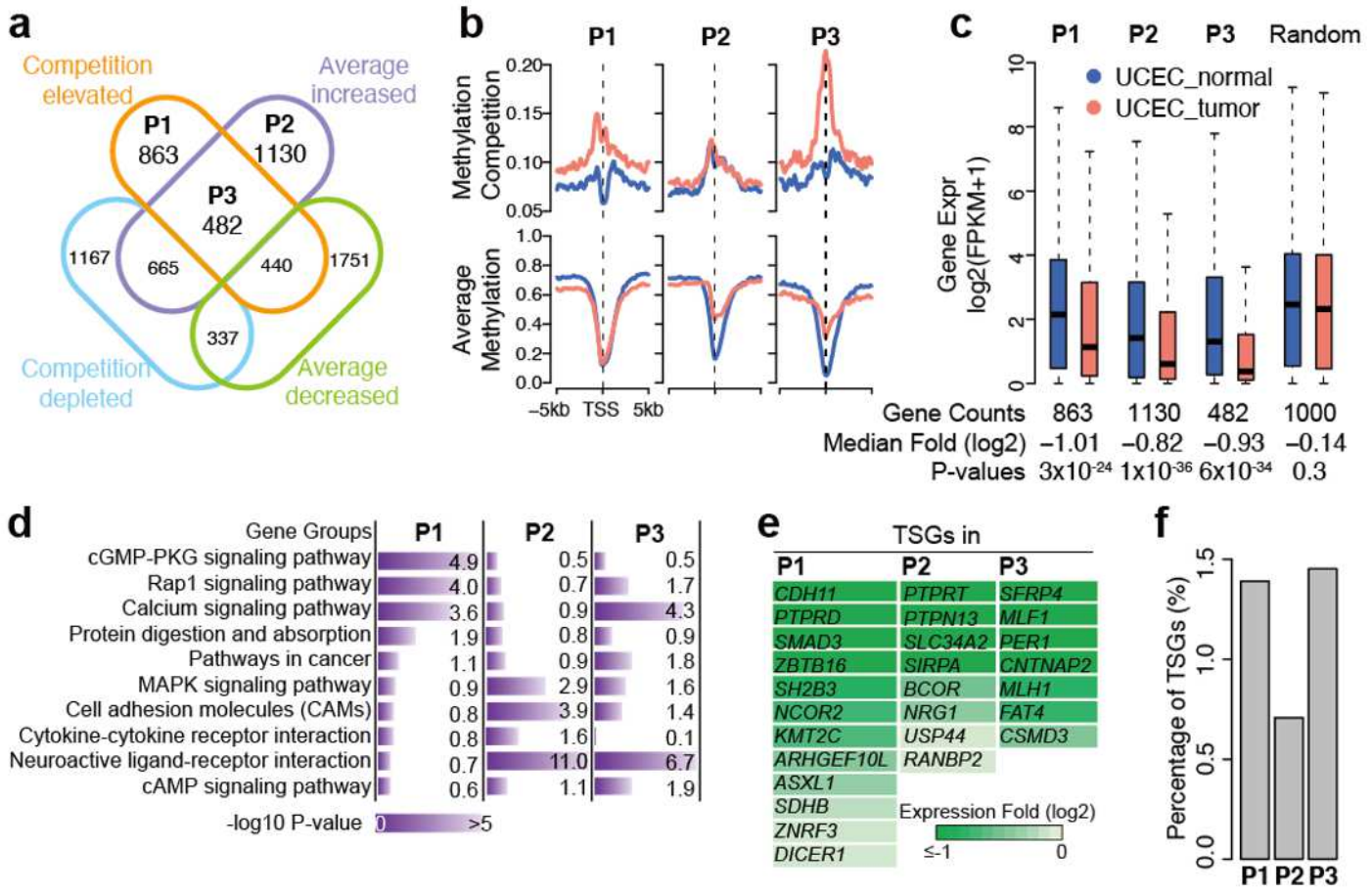

Figure 3

Tumor repressors can be repressed by methylation competition elevation rather than hypermethylation in uterus tumor (a) Venn diagram shows the overlap between promoter methylation competition altered genes and average methylation altered genes in TCGA uterus tumor (UCEC). P1, methylation competition-elevated but average methylation stable genes; P2, hypermethylated but competition-stable genes; P3, competition-elevated and hypermethylated genes. (b) Average profiles of methylation competition (upper panel) and average methylation (lower panel) in TSS regions of P1, P2, and P3 genes. UCEC normal is in blue, and UCEC tumor is in red. (c) Gene expression change of P1, P2, P3, and 1,000 randomly selected genes. The fold changes between median values (log2 scale) are indicated below. The two-tailed Wilcoxon signed-rank test was used for the significance test. The lists of repressed genes in 3 groups are in Supplementary Table 2. (d) Gene ontology analysis on P1, P2, and P3 genes using DAVID online tool. The Benjamini-Hochberg adjusted p-values were given by DAVID. (e) COSMIC tumor suppressor genes which are overlapped with P1, P2, and P3 genes. Gene expression fold changes (log2 scale) are indicated. Darker green means higher repression in the UCEC tumor sample. (f) The percentages of tumor suppressors in P1, P2, and P3 genes.

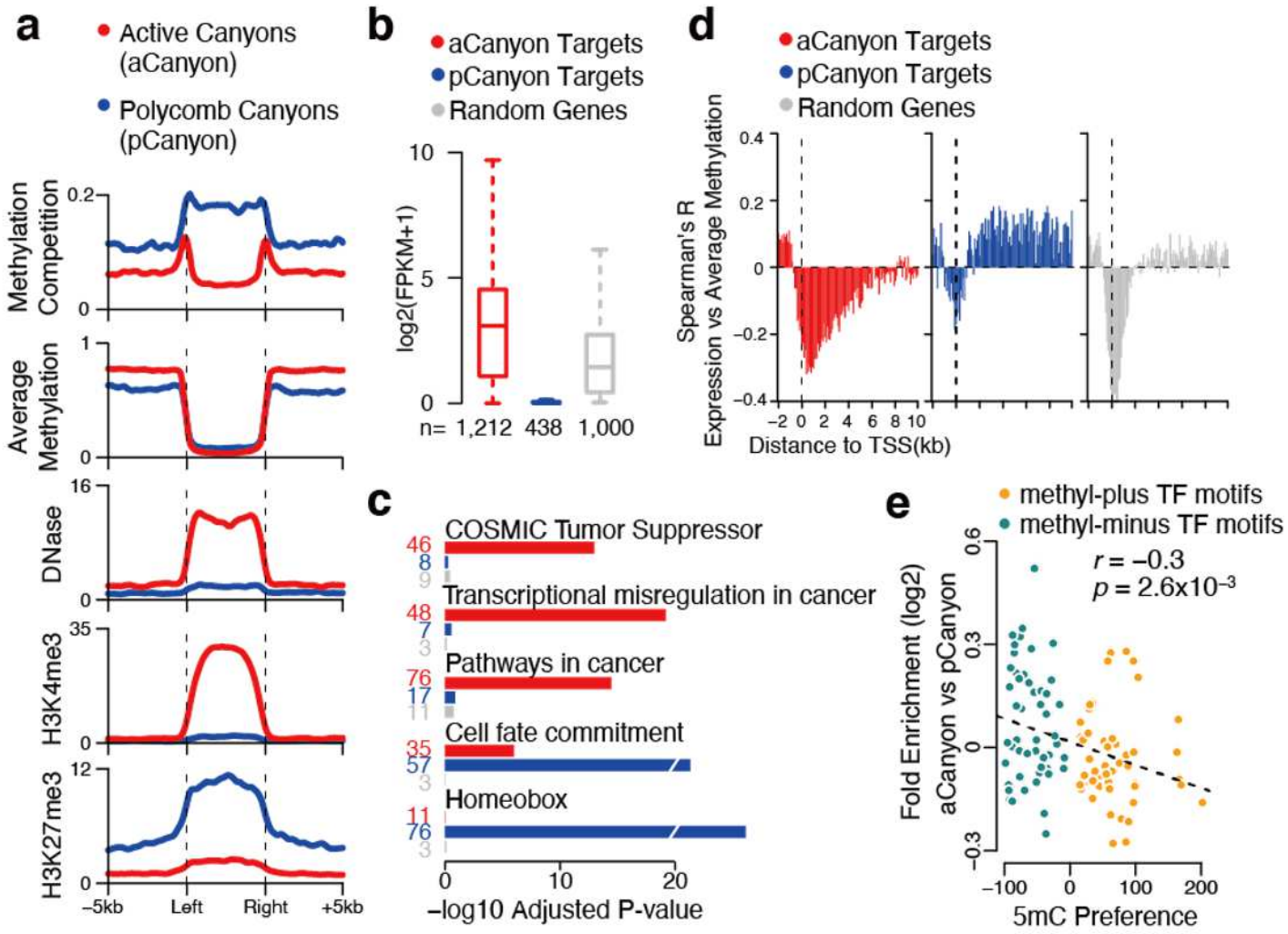

Figure 4

Distinct characteristics of methylation canyons categorized by methylation competition (a) Average profiles of methylation competition, average methylation, DNase I hypersensitive sites, H3K4me3, and H3K27me3 in methylation canyons in CD3+ T-cell. ‘aCanyons’ (red) are low-competition canyons. ‘pCanyons’ (blue) are high-competition canyons. The X-axis indicates the distance to canyon borders. And all canyons are normalized to 5kb length. (b) Expression of aCanyon target genes (red), pCanyon target genes (blue), and 1,000 randomly selected genes (grey). Lists of aCanyons and pCanyons are in Supplementary Table 3, together with their target genes. (c) Functional enrichment analysis of aCanyon target genes, pCanyon target genes, and 1,000 randomly selected genes. Enriched gene counts in each group are indicated on the left side. P-values were measured by two-tailed Fisher’s Exact Test and adjusted by the Benjamini-Hochberg method. (d) Spearman correlation between gene expression and average methylation of 100bp-bin in gene regions. Totally 120 bins from -2kb to 10kb were measured. (e) Relationship between fold enrichment and 5mC preference of TF motifs. Each dot represents a motif. Y-axis indicates the fold change (log2) between enrichment at aCanyon and enrichment at pCanyon of the same motif (see Methods). The X-axis shows the 5mC preference of motifs measured by the SELEX technique. ‘methyl-plus’ TFs prefer to bind methylated sequences, while binding of ‘methyl-minus’ TFs are not favored by 5mC. The list of ‘methyl-plus’ and ‘methyl-minus’ TFs are in Supplementary Table 4. Spearman’s rank correlation was used. P-values were calculated by the two-tailed correlation test for Spearman’s correlation. The linear model was plotted to describe the relationships between variables (indicated by the dashed line).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CAMDASupplementaryV6.docx