Supplementary Document for
Comparative pan-cancer DNA methylation analysis reveals cancer common and specific patterns

Xiaofei Yang¹, Lin Gao¹*, Shihua Zhang²*

¹ School of Computer Science and Technology, Xidian University, Xi’an 710071, Shaanxi, China
² National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

* To whom correspondence should be addressed. Tel/Fax: +86 01 8254 1360; Email: zsh@amss.ac.cn. Correspondence may also be addressed to Lin Gao. Tel/Fax: +86 029 8820 1631; Email: lgao@mail.xidian.edu.cn.

Test the hypermethylation of those three enhancer probes with WGBS data of seven tumor types

Seven of the 15 cancer types have the whole-genome bisulfite sequencing (WGBS) data (BLCA, BRCA, COAD, LUAD, LUSC, READ, and UCEC). However, they only have a limited number of samples for each cancer (BLCA: 6T vs 1N; BRCA: 5T vs 1N; COAD: 2T vs 1N; LUAD: 5T vs 1N; LUSC: 4T vs 1N; READ: 2T vs 1N; UCEC: 5T vs 1N, where T represents tumor samples and N represents normal samples), enabling it hard to get robust statistical test. We validate the methylation levels of those three enhancer probes (cg22620090, cg02829743, and cg02391713) using such WGBS data of these seven tumor types to confirm our finding (Supplementary Figure S8). The results indicate that the methylation level of cg22620090, cg02829743 and cg02391713 is hypermethylated in 4 (BLCA, LUSC, READ, UCEC), 6 (BLCA, BRCA, COAD, LUSC, READ, UCEC) and 4 (BLCA, LUSC, READ, UCEC) tumor types, respectively. These results confirm that these three probes tend to be hypermethylated in many tumor samples.

Pathway enrichment analysis by GOseq

We check the number of probes that located in the promoters or bodies of the PME genes, and find different gene promoters or bodies contain diverse number of probes (from 1 to 1275 in the gene bodies and from 1 to 131 in the gene promoters). Here we adopt the Bioconductor package GOseq [1] to correct the potential bias in the pathway enrichment analysis [2]. We fit the
Probability Weighting Function (PWF) of each genes by nullp function and the parameter bias.data is set as the probe number of each gene. We perform GO BP and KEGG pathway enrichment and merge the results. The terms with FDR (Benjamini-Hochberg corrected) < 0.1 are selected as the significant enriched ones (Supplementary Figure S15), which still confirms that distinct pan-cancer DNA methylation-expression gene groups reveal distinct functional groups. We can also see that many of the enriched terms are closely related in both results (GOseq and g:Profiler). For examples, in HoBP group, most of the enriched results are related with signaling pathway in both, in HBP group, the enriched results are associated with development or differentiation in both results.

**csDMCs tend to be associated with cancer specific functions**

Super-enhancers are a large cluster of transcriptional enhancers which have been proved to play key roles in cell identity of normal cells and regulation of tumor pathogenesis genes [3, 4]. Both hyper- and hypomethylated csDMCs tend to be significantly enriched in cell type-specific super-enhancers (the csDMCs overlapped with such super-enhancers are denoted as csDMC_{se}) (Supplementary Figure S17B and Supplementary Table S13). For example, 44 of 250 BLCA-specific hypermethylated csDMCs are overlapped with bladder-specific super-enhancers \((FC = 3.33, p = 2.6e-12)\). 82 of 295 PAAD-specific hypermethylated csDMCs and 49 of 305 PAAD-specific hypomethylated csDMCs are overlapped with pancreas-specific super-enhancer \((FC = 3.61, p = 6.39e-25 \text{ and } FC = 2.08, p = 8.5e-7)\). Moreover, the csDMC_{se} associated genes tend to be enriched cancer related pathways (Supplementary Figure S17C). For example, some of PAAD hypermethylated csDMC_{se} associated genes (e.g., \textit{CPA1}, \textit{CTRL}, \textit{CELA3A}, \textit{CELA3B}) belong to pancreatic secretion pathway \((FDR = 0.026)\), which plays basic roles in pancreas cells. Some ESCA hypomethylated csDMC_{se} associated genes (e.g., \textit{TNRC6C}, \textit{TNRC6B}, \textit{FURIN}, \textit{NOD2}, \textit{MIB2}, \textit{NEURL1B}) participate in notch signaling pathway \((FDR = 0.037)\), which is related with cell death in esophageal cancer cells [5]. Furthermore, some of csDMC_{se} associated genes are known cancer-related ones with a variety of numbers (ranging from 1 in BLCA hypermethylated to 77 in COAD hypermethylated) (Supplementary Table S14). Most of such genes show significant correlations between their methylation and expression levels (Supplementary Table S14). One hypomethylated csDMC_{se} in ESCA locates in the promoter of \textit{ETS2}, which is an oncogene and overexpressed in esophageal squamous cell carcinoma [6]. Intriguingly, the correlation of the methylation level of this csDMC_{se} with the expression of \textit{ETS2} in ESCA samples (both tumor and normal samples) is significantly negative \((PCC = -0.47, p = 6.9e-12)\) (Supplementary Figure S17D and Supplementary Table S14). Thus, we conjecture that the loss of DNA methylation in promoter results in the expression increase of oncogene \textit{ETS2} in ESCA. These observations suggest that the abnormal methylation of the cell type-specific super-enhancers may contribute to the corresponding cancer genesis.
Supplementary figures

Figure S1
Figure S2
Figure S3
Figure S4
Figure S5
Figure S6
Figure S7
Figure S8
Figure S9
Figure S10
Figure S11
Figure S12
Figure S13
Figure S14
Figure S15
Figure S16
Figure S17
Figure S18
Figure S19
Figure S20
Figure S21
Figure S1. Distributions of DMCs in different cancers. (A) The proportions of DMCs in different genomic regions. (B) The proportions of DMCs in CGI associated regions. (C) The FCs of DMCs in different genomic regions. (D) The FCs of DMCs in CGI associated regions.
Figure S2. Unsupervised hierarchical clustering of common DMCs in 15 cancers using *wald.D2* linkage with Jaccard distance in *hclust* function in R.
Figure S3. The proportions (A) and fold changes (FCs) (B) of hypermethylated PDMCs in different chromosomes. “+” stands for $1e^{-10} \leq FDR < 1e^{-5}$, and “*” stands for $FDR < 1e^{-10}$. 
Figure S4. The proportions (A) and fold changes (FCs) (B) of hypomethylated PDMCs in different chromosomes. “+” stands for $1e$-$10 \leq FDR < 1e$-$5$, and “*” stands for $FDR < 1e$-$10$. 
Figure S5. Distributions of PDMCs in different genomic regions and CGI associated regions. The proportions (A) and fold changes (C) of hypermethylated PDMCs in different regions. The proportions (B) and fold changes (D) of hypomethylated PDMCs in different regions. "+" stands for $1\times10^{-10} \leq FDR < 1\times10^{-5}$, and "***" stands for $FDR < 1\times10^{-10}$.

Figure S6. The statistical numbers of PDMCs hypermethylated in a given number of cancers.
Figure S7. The statistical numbers of PDMCs hypomethylated in a given number of cancers.

Figure S8. The hypermethylation of three enhancer probes (including cg22620090, cg02829743, and cg02391713) with the WGBS data of seven tumor types, including BLCA, BRCA, COAD, LUAD, LUSC, READ and UCEC from TCGA.
Figure S9. Illustration of this hypermethylated enhancer region containing three identified CpG probes which are marked by H3K27Ac and bonded by EZH2. It is plotted using UCSC genome browser. It shows H3K27Ac marker on seven cell lines (including GM12878, H1-hESC, HSMM, HUVEC, K562, NHEK, and NHLF). Different colors represent different cell lines (details showed in UCSC genome browser H3K27Ac Track: http://ucscbrowser.genap.ca/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeRegMarkH3k27ac). The EZH2 bind signal is clustered from eight normal cell lines (including HMEC, HSMM, HSMMtube, HUVEC, NH-A, NHDF-Ad, NHEK, and NHLF).
Figure S10. Annotation of enhancer region we found. (A) The enhancer region in roadmap chromatin state by WashU Epigenome Browser (http://epigenomegateway.wustl.edu/browser/), each row represents one primary somatic cell. The yellow shadow represent the enhancer region chr6: 105400884-105401043. (B) H3K27ac and H3K4me1 annotation of the enhancer region using Roadmap Epigenome Browser (http://epigenomegateway.wustl.edu/browser/roadmap/). The color represents the binding intensity.
Figure S11. The methylation level of two CpG sites (A) cg02829743 and (B) cg02391713 in tumor and normal samples of different cancer types.
Figure S12. The Pearson coefficient correlations (PCC) of methylation of three enhancer region CpG sites with the expression of 10 up- and downstream genes. Only the correlations with statistical significance and |PCC| > 0.1 are shown.
Figure S13. Clustering of motifs enriched in (A) hyper- and (B) hypomethylated PDMCs by "GOSemSim" R package and the pathway enrichment of each cluster.
Figure S14. Heatmap of log2 fold change (obtained from differentially expressed analysis) of target genes of TFs enriched in (A) hypermethylated PDMCs and (B) hypomethylated PDMCs.

Figure S15. Log2 fold change of ASNS, GDF15 (target gene of ATF3) and CCNE1 (target gene of ARNT) across 15 cancer types.
Figure S16. Pathway enrichment of different gene groups by "GOseq". We chose the top 10 enriched pathway in each gene group. HPP: Hypermethylated, Promoter, Positive; HBP: Hypermethylated, Body, Positive; HBN: Hypermethylated, Body, Negative; HoPP: Hypomethylated, Promoter, Positive; HoBP: Hypomethylated, Body, Positive.
Figure S17. Seven survival related hypomethylated PDMCs. (A) The statistical significance of these PDMCs correlating with patient survival time in 15 cancers respectively. The p-value is computed by Wald test. “hr” represents hazard ratio. For convenience, we truncate –log2(p) by 10. (B) Kaplan-Meier survival curve showing overall survival of different cancers. Statistical difference in outcome between high and low index groups is indicated by log-rank test p-value. “+” stands for the censoring samples.
Figure S18. Enrichment analysis of csDMCs. (A) The enrichment of csDMCs associated genes in the known cancer ones. (B) The enrichment of csDMCs with cell type-specific super-enhancers. (C) The pathway enrichment of hypermethylated csDMC\textsubscript{se} genes and hypomethylated csDMC\textsubscript{se} genes, respectively. (D) The correlations between methylation and expression of $ETS2$ in ESCA cancer.
Figure S19. The enrichment of csDMCs in cell type-specific hypoMarks. "*" stands for $1 \times 10^{-5} < p$-value $< 0.05$, and "**" stands for $p$-value $< 1 \times 10^{-5}$.

Figure S20. The distance distributions of the neighboring (A) hypermethylated PDMCs
Figure S21. PDMCs associated genes that showing both significant correlations (A) between methylation and expression and (B) between methylation and patient survival times. In (B), each row represent a PDMC, and the gene is the PDMC associated one. In some case, one gene contain two or more PDMCs. Up plane shows seven hypermethylated PDMC associated genes and down plane shows one hypomethylated PDMC associated gene. Genes in orange means they have been reported as tumor suppressor genes, genes in blue means they have been reported as oncogenes, and genes in black means they have not been determined as tumor suppressor or oncogenes yet.

References