OXFORD

Briefings in Bioinformatics, 18(5), 2017, 761–773

doi: 10.1093/bib/bbw063
Advance Access Publication Date: 18 July 2016
Paper

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

Xiaofei Yang, Lin Gao and Shihua Zhang

Corresponding authors: Shihua Zhang, National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China. Tel.: +86 01 8254 1360; E-mail: zsh@amss.ac.cn; Lin Gao, School of Computer Science and Technology, Xidian University, Xi'an 710071, Shaanxi, China. Tel.: +86 029 8820 1631; E-mail: lgao@mail.xidian.edu.cn

Abstract

Abnormal DNA methylation is an important epigenetic regulator involving tumorigenesis. Deciphering cancer common and specific DNA methylation patterns is essential for us to understand the mechanisms of tumor development. The Cancer Genome Atlas (TCGA) project provides a large number of samples of different cancers that enable a pan-cancer study of DNA methylation possible. Here we investigate cancer common and specific DNA methylation patterns among 5480 DNA methylation profiles of 15 cancer types from TCGA. We first define differentially methylated CpG sites (DMCs) in each cancer and then identify 5450 hyper- and 4433 hypomethylated pan-cancer DMCs (PDMCs). Intriguingly, three adjacent hypermethylated PDMC constitute an enhancer region, which potentially regulates two tumor suppressor genes *BVES* and *PRDM1* negatively. Moreover, we identify six distinct motif clusters, which are enriched in hyper- or hypomethylated PDMCs and are associated with several well-known cancer hallmarks. We also observe that PDMCs relate to distinct transcriptional groups. Additionally, 55 hypermethylated and 7 hypomethylated PDMCs are significantly associated with patient survival. Lastly, we find that cancer-specific DMCs are enriched in known cancer genes and cell-type-specific super-enhancers. In summary, this study provides a comprehensive investigation and reveals meaningful cancer common and specific DNA methylation patterns.

Key words: cancer genomics; DNA-methylation; pan-cancer study; bioinformatics

Introduction

DNA methylation is a well characterized epigenetic marker that is closely associated with oncogenesis [1]. Differential methylation has been identified in many cancers, indicating their potential roles in diverse processes of cancers. Herman *et al.* [2] revealed that *VHL* (Von Hippel-Lindau Tumor Suppressor, E3 Ubiquitin Protein Ligase) is epigenetically silenced owing to hypermethylation of a CpG island (CGI) in the 5′ region in renal cell cancers. Fleischer *et al.* [3] identified 18 CpG probes associated with survival of breast cancer patients through comparative analysis of DNA methylation changes between 46 normal and 239 ductal carcinoma *in situ* samples. Hinoue *et al.* [4] discovered four distinct subgroups in colorectal cancer by analyzing a large-scale genome-wide DNA methylation profiles. Collectively, these studies reveal diverse functions of DNA methylation alterations in individual tumor types.

With the rapid development of biological technology, large-scale methylomes across different tissues, cell lines and samples have been generated by different sequencing or microarray techniques, such as whole-genome bisulfite sequencing (WGBS), reduced representation bisulfite sequencing (RRBS) and Illumina Infinium HumanMethylation450 BeadChips (HM450) [5]. Based on these methylation profiles, researchers have identified novel CGIs [6], detected cell-type-specific DNA methylation marks [7] and detected DNA methylation patterns in and across diverse normal tissues or cancer types (reviewed in [8, 9]). For instance, Yang *et al.* [10] constructed a weighted co-methylation network based on methylomes of 54 cell lines and detected seven distinct biological relevant co-methylation

Xiaofei Yang is a PhD candidate of Xidian University. His research interests include bioinformatics, epigenomics and data mining.
Lin Gao is a Professor of Xidian University. Her research interests include bioinformatics, data mining and machine learning.
Shihua Zhang is an Associate Professor at National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences. His research interests are mainly in bioinformatics, cancer genomics, epigenomics, data mining and machine learning.
Submitted: 26 February 2016; **Received (in revised form):** 2 June 2016

modules. Varley *et al.* [1] analyzed dynamic DNA methylation patterns across a large-scale DNA methylation profiles of 82 human cell lines and tissues and revealed a hypermethylation pattern in cancer cell lines. Gevaert *et al.* [11] developed MethylMix and identified 10 pan-cancer clusters, which reveal new similarities across different cancer types. However, there are still some open questions that remain to be addressed about diverse cancers. Recently, Witte *et al.* [8] surveyed the DNA methylation patterns discovered in individual cancers, and stressed that identification of common DNA methylation patterns will provide new prognostic biomarkers and novel insights into the mechanisms of DNA methylation in cancer development.

Taking the advantage of large-scale DNA methylation data produced by The Cancer Genome Atlas (TCGA) project, we investigate pan-cancer wide DNA methylation patterns among 5480 samples across 15 different cancer types in this study. First, we detect differentially methylated CpG sites (DMCs) in each cancer type and obtain 5450 hyper- and 4433 hypomethylated pan-cancer wide DMCs (PDMCs). Interestingly, we find an enhancer region hypermethylated in >11 cancer types regulates two tumor-suppressor genes *BVES* and *PRDM1* negatively. These two genes have been proved to act as tumor suppressor roles in many cancers such as lung cancer [12], gastric cancer [13] and colon cancer [14, 15]. Therefore, this enhancer region may be a potential epigenetically therapeutic target for many cancers. We further find that these PDMCs show distinct binding characteristics, which are associated with multiple cancer hallmarks such as evading growth suppressors, immune destruction and sustaining proliferative signaling. We observe eight distinct pan-cancer wide methylation-expression gene groups relating to known cancer pathways. We also discover 55 hyper- and 7 hypomethylated PDMCs associated with patient survivals. Lastly, we determine a set of cancer-specific DMCs (csDMCs), which are enriched in known cancer genes, cell-type-specific super-enhancers and methylation marks, indicating their underlying implications to individual cancers.

Materials and methods

Cancers and samples

In this study, we select 15 cancers with at least five normal samples in each. These cancers include bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), esophageal carcinoma (ESCA), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), pancreatic adenocarcinoma (PAAD), prostate adenocarcinoma (PRAD), rectum adenocarcinoma (READ), thyroid carcinoma (THCA) and uterine corpus endometrial carcinoma (UCEC). Totally, there are 5480 samples including 4817 tumor samples and 663 normal samples (Supplementary Table S1).

DNA methylation data

The DNA methylation data (level three) generated using HM450 platform [16] are downloaded from TCGA data portal (https://tcga-data.nci.nih.gov/tcga/) for all samples. The methylation level of each probe is measured as beta value, which is calculated as the ratio of methylated signal to the sum of methylated and unmethylated signal. The range of beta value is from 0 (unmethylated) to 1 (completely methylated), and the missing value is represented as 'NA' which is due to masking of CpGs owing to single-nucleotide polymorphisms (SNPs) with high minor allele frequency (MAF) within 10 bp of the targeted CpGs or substantial overlap between probe sequences and repetitive elements. In total, there are 482 421 CpG probes in the HM450 Beadarray, including 88 445 probes with 'NA' values among >50% samples in at least one cancer. We remove these 88 445 probes and obtain 393 976 CpG probes finally. We impute the remaining 'NA' values using 10-nearest neighbors imputation procedure [3] with the '*knnimpute*' function in MATLAB. The HM450 DNA methylation data of normal and cancer cell lines and the CpG probe annotation file are downloaded from ENCODE project (http://genome.ucsc.edu/ENCODE/downloads.html) [17]. The HM450 data of tumor samples of glioblastoma (GBM) are downloaded from TCGA and the data of normal sample are downloaded from GEO (GSE41826, which contains 58 normal glial cell samples) [18]. The normal data are generated at the same organization via the same pipeline. The HM450 data of tumor and normal samples of acute myeloid leukemia (AML) are also downloaded from GEO (GSE63409, which contains 30 normal bone marrow samples and 44 AML tumor samples) [19]. The information in the CpG probe annotation file includes the corresponding gene, genomic region [regions from 1500 bp upstream to the transcription start site (TSS1500), regions from 200 bp upstream to the transcription start site (TSS200), 5′UTR, 1stExon, Body and 3′UTR] and CGI-associated regions [Island, Shore (up or down 2 kb from CGI), Shelf (up or down 2–4 kb from CGI) and Open sea (> 4 kb from CGI)].

Gene expression data

The TCGA RNASeqV2 data (level three) generated from IlluminaHiSeq_RNASeqV2 platform are downloaded from TCGA data portal (https://tcga-data.nci.nih.gov/tcga/). We obtain the expression data of '*.rsem.genes.normalized_results*' format, in which the expression of genes is quantitated with RSEM (RNA-seq by Expectation Maximization) and normalized by upper quartile normalization.

Overall survival time data

The clinical data of *Biotab* format are downloaded from TCGA data portal (https://tcga-data.nci.nih.gov/tcga/) for survival analysis.

Known cancer genes

We integrate three cancer gene databases including NCI cancer gene index (https://wiki.nci.nih.gov/x/hC5yAQ, accessed on 4 September 2015), COSMIC somatic mutation catalog [20] and CCGD mouse cancer driver genes [21] to generate the known cancer gene set.

Cell-type-specific super-enhancers

The cell-type-specific super-enhancers are obtained from the data set of a recent study [22], which provides a catalog of super-enhancers of 86 human cell lines and tissue samples. We obtain the super-enhancers of five normal tissues (bladder, colon, esophagus, lung and pancreas) that match to six cancer types (BLCA, COAD, ESCA, LUAD, LUSC and PAAD) we used.

Cell-type-specific hypomethylation marks

The cell-type-specific hypomethylation marks (hypoMarks) are downloaded from http://fame.edbc.org/methymark, which is

detected by Liu *et al.* [7] through an entropy-based method. We obtain methylation marks of six tissues (including breast, colon, esophagus, liver, lung and pancreas) that match to seven cancer types (BRCA, COAD, ESCA, LIHC, LUAD, LUSC and PAAD) we used.

Identification of differentially methylated CpG probes

We use the R package '*limma*' [23] designed based on a linear model to identify DMCs between tumor and normal samples with false discovery rate (FDR) ≤ 0.01 and the difference of median DNA methylation level between normal and tumor samples ≥ 0.2. If the methylation level of a DMC is higher in tumor samples than that in normal samples, we define it as hypermethylated one, otherwise hypomethylated one. If a DMC is differentially methylated in at least eight cancers with consistent trend (e.g. hypermethylated or hypomethylated), we consider it as a PDMC. If a DMC is differentially methylated in only one cancer, we define it as a csDMC.

Differential gene expression analysis

We detect the differentially expressed genes by Bioconductor package '*DESeq2*' [24]. We select the genes with adjusted *p-value* < 0.01 and *log2(FC)* > 1 as the over-expressed ones and genes with adjusted *p-value* < 0.01 and *log2(FC)* < -1 as the down-expressed ones.

Motif enrichment analysis

We adopt the HOMER software (http://homer.salk.edu/homer/) [25] to detect sequence motifs that are enriched in the hyper- and hypomethylated PDMCs. The background sequences are generated from the 393 976 CpG probes.

Pathway enrichment analysis

To evaluate the functional relevance of DMCs, we perform pathway enrichment analysis for genes associated with DMCs via *g:Profiler* web server (http://biit.cs.ut.ee/gprofiler/) [26]. The pathway database we used in *g:Profiler* includes Gene Ontology (GO) biological processes, Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome. The significant enrichment pathways are obtained with FDR < 0.05.

Determination of the hyper- or hypomethylation states of different chromosomes or functional regions

Given a cancer type, we compute the methylation difference of a chromosome or functional region as the difference of median values of DMCs in this region between normal and tumor samples. For example, we denote the beta value of DMC c in functional genomic region r in sample s of cancer i as $M_i^{c,r,s}$, the methylation difference is computed as $DM_i^r = \text{median}\left(M_{i,N}^{1,r}, \cdots, M_{i,N}^{p,r}\right) - \text{median}\left(M_{i,T}^{1,r}, \cdots, M_{i,T}^{p,r}\right)$, where p is the number of probes in region r, $M_{i,N}^{c,r}$ is the median methylation value of probe c in region r of all normal samples, $M_{i,T}^{c,r}$ is the median methylation value of probe c in region r of all tumor samples. The significance of the difference is tested by Wilcoxon rank-sum test and FDR is obtained by Benjamini–Hochberg correction. If $DM_i^r < 0$ and FDR < 0.05, region r is hypermethylated in tumor, while if $DM_i^r > 0$ and FDR < 0.05, region r is hypomethylated in tumor.

The common DMCs of different cancers

We adopt the Jaccard coefficient similarity to measure the similarities of cancers based on the number of their common DMCs:

$$J_{ij} = \frac{|DMC_i \cap DMC_j|}{|DMC_i \cup DMC_j|},$$

where DMC_i and DMC_j are the DMCs in cancer i and j, respectively, and J_{ij} is the Jaccard coefficient similarity between cancer i and cancer j. We perform an unsupervised hierarchical clustering on different cancers based on Jaccard distance ($d_{J_{ij}} = 1 - J_{ij}$) to see whether the number of common DMCs can reflect some biological implications or not.

Correlation between DNA methylation and gene expression

Because DNA methylations in promoter and gene body regions have different impacts on gene expression [1, 10], we divide hyper- and hypomethylated PDMCs into promoter (or gene body) associated ones based on their position in promoter (or gene body) region or not. If a gene promoter or body region contains two or more PDMCs, the methylation profile is represented as the scores on the first principle component of the methylation profiles of these PDMCs. We compute Pearson coefficient correlations (PCCs) between the methylation profiles and the expression profiles of PDMC-associated genes for each cancer using the matched RNA-seq and HM450 data. Finally, we select the genes that show significant correlations ($|PCC| > 0.2$, FDR < 0.05) in at least eight cancer types.

Determination of target genes of a specific hypermethylated enhancer region

It is known that enhancers can not only regulate the closest genes, but also further ones in either orientation [27]. To determine target genes of this enhancer region we identified, we compute PCCs between methylation of each enhancer probes with expression of 10 genes upstream and 10 genes downstream from this enhancer region in each cancer type because most of the enhancer target genes are within this distance [28]. We define a gene with significantly positive or negative correlation (FDR < 0.05, $|PCC| > 0.1$) in at least eight cancers as a potential target gene of one enhancer probe. If a gene is the potential target gene of $\lceil p/2 \rceil$ probes in the enhancer region (p is the number of probes in this region), it is considered as the target gene of this enhancer region.

Survival analysis

For each cancer, we use the function *coxph* (R package *survival*) [29] to implement the Cox proportional hazard model to analyze the association of methylation profile of each PDMC with patient survival in each cancer. We consider PDMCs with $P < 0.05$ (Wald test) as the survival correlated PDMCs. We consider PDMCs that are survival related and have the consistent risk tendency (hazard ratio > 1 or hazard ratio < 1) in at least five cancers as pan-cancer wide survival correlated PDMCs.

For each cancer, we adopt the following prognostic index (PI) [3] to generate high- and low-risk patient groups in each cancer:

$$PI_i = \sum_{c=1}^{n} \beta_c m_{ci},$$

where n is the number of survival correlated PDMCs, β_c is the

regression coefficient of Cox proportional hazard model for PDMC c, m_{ci} is the methylation level of PDMC c in patient i. Patients are divided into high- and low-risk groups based on the median PI. We adopt Kaplan-Meier estimator and log-rank test with the functions *survfit* and *survdiff* (R package *survival*) to test the survival difference between two patient groups.

Enrichment analysis of the csDMCs in known cancer genes, super-enhancers and methylation marks

We adopt fold change (FC) and hypergeometric test to measure the overlapping significance between csDMCs and known cancer genes, cell-type-specific super-enhancers and cell-type-specific hypoMarks.

Results

Characterization of differentially methylated CpG sites in 15 cancers

We obtain a set of differentially methylated CpG sites (DMCs) for each of the 15 cancers with different numbers (ranging from 3722 in THCA to 57 290 in UCEC). Most of these DMCs locate in non-promoter regions (gene body, 3′UTR and intergenic region; Supplementary Figure S1A and C) and non-CGI regions (shore, shelf and open sea; Supplementary Figure S1B and D) in each cancer, which confirms that DNA methylation level in promoter and CGI regions tend to be stable [10]. Moreover, DMCs are enriched in intergenic and open sea regions in most of the cancers, indicating that these regions show potential cancer-associated mechanisms.

We note that biologically similar cancers share more common DMCs (Supplementary Figure S2). For examples, the smoking-related, upper aerodigestive tract cancers (LUSC, HNSC and LUAD) are clustered together and share a large number of DMCs. READ and COAD are clustered together with the largest similarity ($J = 0.53$), confirming that they share more similar DNA methylation mechanisms [30].

The methylation states of most cancers are diverse in different chromosomes and functional regions including TSS1500, TSS200, 5′UTR, 1stExon, body, 3′UTR and intergenic regions (Figure 1). For instance, HNSC is significantly hypermethylated in chromosome 4 (FDR = 7.18e-14, Wilcoxon rank-sum test) while significantly hypomethylated in chromosome 8 (FDR = 2.35e-16) (Figure 1A). COAD is significantly hypermethylated in promoter regions (include TSS1500, TSS200, 5′UTR and 1stExon) (FDR < 1.98e-168) while significantly hypomethylated in 3′UTR (FDR = 4.46e-23) (Figure 1B). In addition, some cancers (e.g. KIRP and PRAD) are significantly hypermethylated in all functional regions (FDR < 2.2e-9), while some cancers (e.g. LIHC, BLCA and THCA) are significantly hypomethylated in all functional regions (FDR < 0.0045) (Figure 1A and B). We can clearly see that cancers are divided into two classes (globally hypermethylation versus hypomethylation) based on the methylation difference in different chromosomes and different functional regions, respectively ('Materials and Methods' section) (Figure 1). This suggests that the roles of methylation in cancer are of two types, and hyper- or hypomethylated events might be biomarkers of cancers as suggested by individual cancer studies like in prostate carcinoma [31] and bladder cancer [32]. Moreover, we perform differential gene expression analysis on each cancer type, and count the over- and down-expressed genes (Supplementary Table S2). The results show that most (six of seven) of the hypomethylated cancers have more over-expressed genes than down-expressed ones and half (four of eight) of hypermethylated cancers have more down-expressed genes than over-expressed ones, suggesting that the global differentially DNA methylation patterns are relevant to the differentially expressed gene patterns.

Global characterization of PDMCs

We identify 5450 hypermethylated PDMCs and 4433 hypomethylated PDMCs, and find that the hypermethylated PDMCs are significantly enriched in chromosomes 5 and 18, and depleted in chromosomes 16, 17 and 22 (Supplementary Figure S3), while the hypomethylated PDMCs are significantly enriched in chromosomes 8 and 20 and depleted in chromosome 17 (Supplementary Figure S4). Similar with the distribution of DMCs, we also see that hyper- or hypomethylated PDMCs are significantly enriched in intergenic regions (Supplementary Figure S5). Interestingly, the hypermethylated PDMCs are significantly enriched in CGI while depleted in non-CGI regions; however, the hypomethylated PDMCs are significantly enriched in non-CGI regions while depleted in CGI regions. These results confirm that CGI regions tend to be hypermethylated in cancers [33].

A hypermethylated enhancer region occurred in at least 11 cancers is negatively correlated with two tumor suppressor genes

PDMCs that hypermethylated or hypomethylated in all or most cancer types are of particular interest; 1024 of 5450 hypermethylated and 549 of 4433 hypomethylated PDMCs occurred in >11 cancer types (Supplementary Figures S6 and S7). Among them, 11 PDMCs occurred in 14 cancers (10 hyper- and one hypomethylated) located in promoter or gene body regions of seven genes, which are known to be associated with tumor progression (Supplementary Table S3). Four of the 10 hypermethylated ones locate in the promoter of *DRD5*, which shows distinct implications with various cancers. For instance, it is hypermethylated in most subtypes of breast cancer [34]; it acts as candidate drug target in acute lymphoblastic leukemia [35]; and it is significantly downregulated in gastric cancer [36]. Two of the 10 hypermethylated ones locate in the promoter of *RUNX3*, which has been reported to play tumor suppressor roles in breast cancer [37], gastric cancer [38] and prostate cancer [39].

Intriguingly, one probe, cg22620090, is hypermethylated in all 15 cancers (Figure 2A). We also validate the hypermethylation of this probe in tumor samples of GBM, AML, tumor cell lines in ENCODE (Figure 2B) and the WGBS data in seven cancer types (Supplementary Figure S8). cg22620090 locates in chromosome 6: 105400993–105401043 and is marked by an active enhancer marker H3K27ac [40] (Supplementary Figures S9 and S10B). In a recent study, Yao *et al.* [27] also identified the hyper- and hypomethylated enhancer CpG probes in HM450 Beadchip from cancer methylomes, and indicated that cg22620090 is a hypermethylated enhancer probe in eight cancers. Additionally, two other probes, cg02829743 and cg02391713, which locate within the 200 bp upstream of cg22620090 and are hypermethylated in 11 cancers and 13 cancers, respectively (Supplementary Figures S8 and S11), are also marked by H3K27ac (Supplementary Figures S9 and S10B) and have been determined as hypermethylated enhancer probes [27]. Moreover, we check the region in roadmap chromatin states [41] (Supplementary Figure S10A), which indicates that the region we found is annotated as *TSS*, *enhancer* and *Polycomb* associated states in different normal primary somatic cells. Together, these three probes constitute a hypermethylated enhancer region. We find that two genes (*BVES*, the downstream second closest gene, and *PRDM1*,

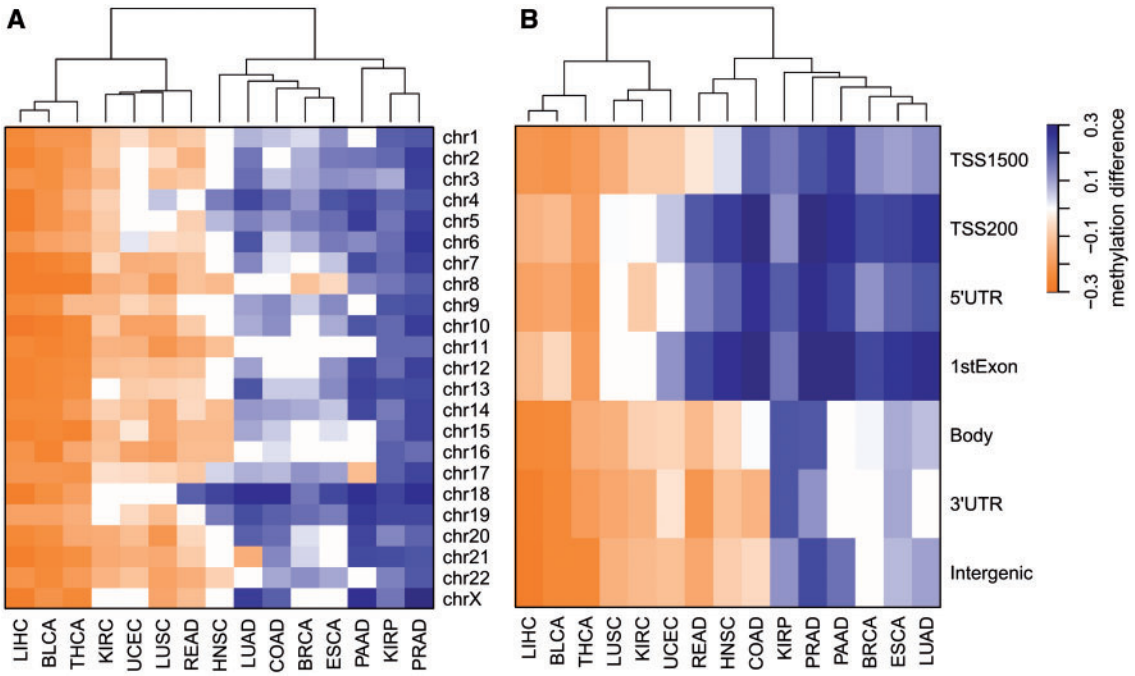

Figure 1. Global methylation states of (**A**) different chromosomes and (**B**) different genomic regions in 15 cancers. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

the downstream fifth closest gene) are potential target ones of this region ('Materials and Methods' section) (Figure 2C and Supplementary Figure S12). We validate the interactions between this enhancer region and *BVES* as well as *PRDM1* by Hi-C data in IMR90 fetal lung fibroblasts [42] (download from http://chromosome.sdsc.edu/), the normalized interaction scores between the enhancer region and *BVES* and *PRDM1* are 13.6276 (top 0.18%) and 2.1864 (top 1.58%), respectively, which confirm *BVES* and *PRDM1* are potential target genes of this enhancer region. Clearly, their expressions are negatively correlated with the methylations of probes in this enhancer region ($|PCC| > 0.1$, $FDR < 0.05$) (Figure 2C). It has been showed that *BVES* acts as a tumor suppressor in many cancers, such as non-small-cell lung cancer [12], gastric cancer [13], hepatocellular carcinoma [43] and colorectal cancer [14]. *PRDM1* has also been shown to act as tumor suppressor in diffuse large B cell lymphoma [44] and colon cancer [15]. These suggest that the hypermethylated enhancer regions might result in the decreasing expression of *BVES* and *PRDM1*, which contribute to the loss of tumor suppressor functions in tumors. Therefore, this enhancer region may be a potential epigenetically therapeutic target for many cancers.

It is known that the methylation changes of an enhancer region can be owing to gain or loss of some transcription factor (TF) bindings [27]. To explore the reason of the hypermethylation of this enhancer region (containing cg22620090, cg02829743 and cg02391713), we detect the TFs that bind to this enhancer region by UCSC genome browser [45] and find that *EZH2* binds to this region in multiple normal cell lines (Supplementary Figure S9). We conjecture that the hypermethylation of this enhancer region in tumor may be resulted from the loss of *EZH2* binding, which is consistent with a previous observation that hypermethylated regions in cancer cell lines lacks *EZH2* bindings [1]. Together, these results support a simple model (Figure 2D) that this enhancer region is hypomethylated and bounded by *EZH2* in normal cells and it activates the tumor suppressor genes *BVES* and *PRDM1*; while in cancer cells, this region is hypermethylated owing to loss of *EZH2* binding and it turns to repress the expression of *BVES* and *PRDM1*.

TFs binding to PDMCs are associated with various cancer hallmarks

We explore the TF binding motifs enriched in hyper- and hypomethylated PDMCs by HOMER software [25], and identify 27 and 31 motifs in hyper- and hypomethylated PDMCs, respectively (Figure 3 and Supplementary Table S4). The motif of *ZFP161* is the most significantly enriched one ($P = 1.0e\text{-}51$) in hypermethylated PDMCs (Figure 3A and Supplementary Table S4) and the motif of *ATF3* is the most significantly enriched one ($P = 1.0e\text{-}118$) in hypomethylated PDMCs (Figure 3D and Supplementary Table S4). *ZFP161* has been shown to act as an activator of the transportation of dopamine [46], which can block the angiogenesis process in tumor [47] and was considered as an option for cancer treatment [48]. *ATF3* has been shown to act as an oncogene in breast cancer [49, 50] and non-small-cell lung cancer [51]. These observations indicate that the TFs tending to bind PDMCs play roles in tumor development.

We cluster the TFs based on the sequencing similarities by STAMP web server [52] and GO semantic similarities by 'GOSemSim' R package [53], respectively. We obtain 12 TF clusters (six hyper-type and six hypo-type) (Figure 3B and E) by STAMP, and eight TF clusters (three hyper-type and five hypo-type) by GOSemSim (Supplementary Figure S13), respectively. Most of these TF clusters are enriched in distinct cancer-associated pathways (Figure 3B and E, and Supplementary Figure S13). For example, cluster_VI in STAMP hyper-type (including *NKX2-5*, *E2F4*, *SMAD3* and *ZFP161*) is significantly enriched in 'cell cycle pathway' ($FDR = 5.5e\text{-}3$), which relates to the sustaining proliferative signaling cancer hallmark, and 'TFG-beta signaling pathway' ($FDR = 2.5e\text{-}3$), which is associated with evading growth suppressors cancer hallmark [54] (Figure 3B). Cluster_IV in STAMP hypo-type (including *MECOM*, *RUNX1*, *RUNX2* and *ESRRA*) is significantly enriched in 'immune system-associated pathways' ($FDR < 0.013$), which relates to evading immune destruction hallmark (Figure 3E). Cluster_III in

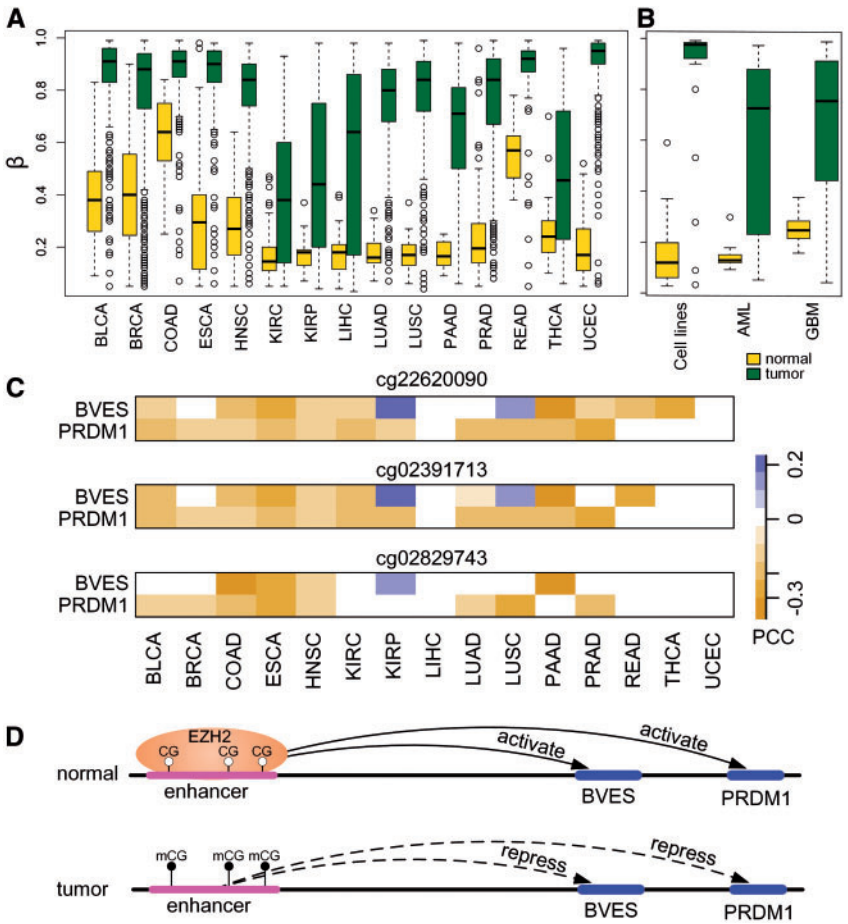

Figure 2. A typical hypermethylated enhancer region. (**A**) The methylation level difference of cg22620090 probe in 15 cancer types between normal and tumor samples. (**B**) Confirmation of the methylation level difference of cg22620090 using normal and tumor cell lines in ENCODE, normal and tumor samples in GBM and AML. (**C**) The PCCs of the expression of potential target genes with the methylation of three probes in this enhancer region among 15 cancer types. (**D**) A model of this enhancer region regulating its target genes. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

GOSemSim hyper-type (including *ESRRA*, *E2F4*, *CEBP*, *NFY*, *NRF1*, *E2F2*, *MZF1*, *MTF1*, *ZFP161*) is significantly enriched in 'PPARA activates gene expression' (FDR = 3.39e-3), which is associated with cancer cell growth [55] and 'Cyclin D-associated events in G1' (FDR = 0.0287), which is associated with 'cell cycle' (Supplementary Figure S13A). Interestingly, we find some clusters from different methods have significantly overlapped TFs (Figure 3C and F), these clusters share the same enriched pathways and the overlapped TFs have been proved to cooperate in some cancer-associated process or interact with each other. For instance, for the hypo-type, *MAX* and *MYCN* are clustered together in both methods ($P = 0.03$, hypergeometric test), the corresponding clusters (cluster_II from STMAP and cluster_IV from GOSemSim) are enriched in 'transcriptional misregulation in cancer' (FDR = 0.012 versus FDR = 0.011) (Figure 3E and Supplementary Figure S13B). It has been shown that *MAX* and *MYCN* are oncogenes and physically interact with each other to form *MYC:MAX* dimers that promote cell proliferation and programmed cell apoptosis in cancer cells [56]. All of these indicate that hyper- or hypomethylated PDMCs are involved in diverse cancer-associated pathways, and play important roles in tumorgenesis.

Furthermore, we study the associations of the TFs enriched in hyper- or hypomethylated PDMCs and the expressions of their target genes (the TF target genes were downloaded from TRRUST [57]). We find that some of the target genes of TFs enriched in both hyper- and hypomethylated PDMCs are over-expressed or down-expressed in multiple cancer types (Supplementary Figure S14). More importantly, some of these target genes play important roles in cancer development. For example, *ATF3* is the most significantly hypomethylated PDMCs enriched TF ($P = 1.0e-118$), which activates two genes, *GDF15* and *ASNS*. These two genes are significantly over-expressed in at least eight cancer types (Supplementary Figure S15) and play oncogenic roles in cancer progression [58, 59]. This may suggest that hypomethylation of *ATF3* allows its binding to activate these two genes to contribute to tumorigenesis. *ARNT*, which represses one of its target gene *CCNE1*, is one of the TFs enriched in hypermethylated PDMCs ($P = 1.0e-20$). Interestingly, *CCNE1* is also significantly over-expressed in 11 cancer types (Supplementary Figure S15) and plays oncogenic roles in many tumor types via regulation of transition from G1 to S phase in cell cycle [60]. This indicates that hypermethylation blocks binding of *ARNT* and results in over-expression of *CCNE1* in tumor samples. These results demonstrate that the roles of hyper- or hypomethylation of TF binding sites contribute to cancer development via regulating the expression of their target genes.

Distinct pan-cancer DNA methylation-expression gene groups reveal distinct functional groups

We explore the correlation of methylation and expression profiles of PDMC-associated genes and find 812 genes with distinct

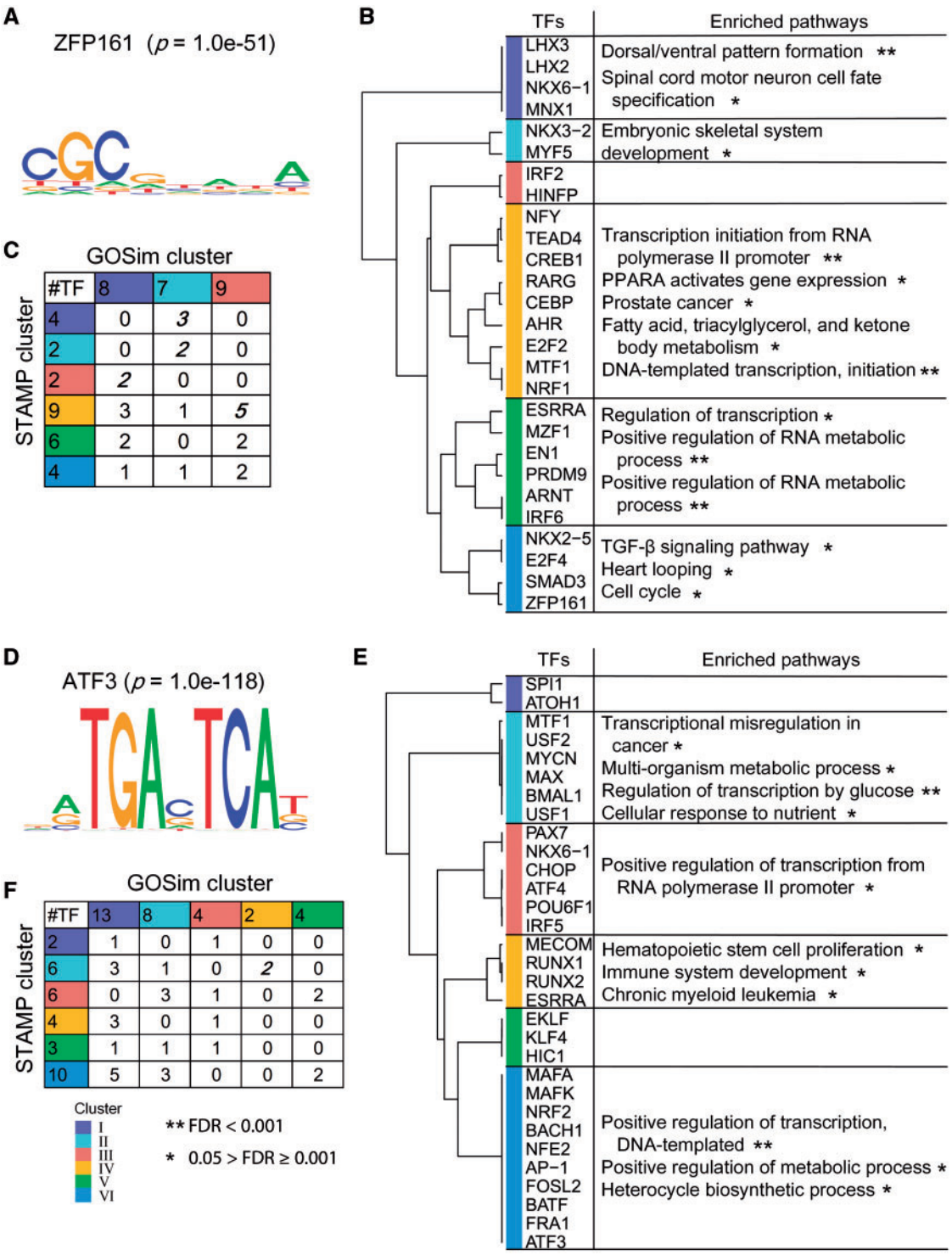

Figure 3. TF binding motifs enriched in hyper- and hypomethylated PDMCs. (**A, D**) the most significantly enriched motifs in (**A**) hyper- and (**D**) hypomethylated PDMCs. (**B, E**) the clustering of TFs enriched in (**B**) hyper- and (**E**) hypomethylated PDMCs by STAMP web server and the pathway enrichment of each cluster. (**C, F**) the number of overlapped TFs in the clusters of motifs enriched in (**C**) hyper- and (**F**) hypomethylated PDMCs by the two different methods. The numbers in the first column and the first row indicate the number of TFs in the corresponding clusters. The bold and italic numbers in (**C**) and (**F**) indicate hypergeometric test $P < 0.1$. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

positive or negative correlations in at least eight cancers, which are denoted as pan-cancer methylation-expression (PME) genes (Figure 4A and Supplementary Table S5). Based on hyper- or hypomethylation, the location of PDMCs in promoter or gene body, and the positive or negative correlation, we obtain eight different PME gene groups (Figure 4A and Supplementary Table S5), which are enriched in distinct pathways (Figure 4B). Distinctly, several groups are involved in cancer-associated pathways. For example, HPN group is significantly enriched in 'Transcriptional misregulation in cancer' (FDR = 4.84e-3). HoPN

group is significantly enriched in 'p53 signaling pathway' (FDR = 3.45e-3), which can control cell cycle and block cancer progression [61]. HoPN group is also significantly enriched in 'TRAIL-activated apoptotic signaling pathway' (FDR = 5.87e-3), which is associated with initiating cell death in various tumor cells [62]. HoBP group is significantly enriched in 'Rap1 signaling pathway' (FDR = 1.48e-6), which is associated with the regulation of myelopoiesis and delivers the oncogenic signals in cancer cells [63]. In summary, the DNA methylation- and expression-correlated genes demonstrate distinct cancer relevance, which should be valuable to decipher their underlying implications. In addition, we apply the Bioconductor package '*GOseq*' [64, 65] to correct the potential pathway enrichment bias (induced by the diverse number of probes in PME genes), and the results also confirm that distinct pan-cancer DNA methylation-expression gene groups reveal distinct functional groups (Supplementary Materials and Supplementary Figure S16).

Pan-cancer survival analysis reveals survival-related PDMCs

We relate the methylation level of PDMCs to patient overall survival in each cancer and obtain 55 hypermethylated PDMCs and seven hypomethylated PDMCs, which are significantly prognostic ones in more than five cancers (Figure 5A and Supplementary Figure S17A and Table S6). Moreover, cancers can be clustered into three clusters based on the Wald test *P*-value of 55 hazardous (hazard ratio $>$ 1) hypermethylated PDMCs (Figure 5A), suggesting that DNA methylation has different effects on patient survival times in different cancers.

Of the 55 hazardous hypermethylated PDMCs, 35 are in CGI, and 27 are associated with 24 genes, including 21 protein-coding genes and 3 noncoding ones (Figure 5A). Seventeen of the 21 hypermethylated PDMC-associated genes have been reported to be associated with cancers (Supplementary Table S7). For example, it has been shown that the hypermethylation of *NEUROG1*, *PAX6* and *TBX5* is associated with poor survival in colorectal cancer [66, 67], gastric cancer [68] and colon cancer [69], respectively. We also see that these genes are significantly enriched in cancer-associated pathways (Figure 5B) such as 'blood vessel development' (FDR = 0.0094), which is associated with 'inducing angiogenesis' cancer hallmark [54], 'tumor necrosis factor production' (FDR = 0.0054) and 'regulation of execution phase of apoptosis' (FDR = 4.74e-4). Of the three noncoding genes, *LOC645323* is associated with colorectal cancer [70] and hypermethylated in breast cancers [71]. Moreover, it also plays hub roles in cancer gene co-expression network [72]. Of the seven survival-associated hypomethylated PDMCs, five locate in open sea region, and three are associated with three genes (Supplementary Figure S17A). One of the three is *DCP2*, coding an mRNA-decapping enzyme, which is an important component of *HIT* superfamily contributing to bioenergetics of cancer cells [73]. Another one is *TMEM132C*, coding a transmembrane protein, which plays key roles in cancer cells [74].

We divide the patients into two groups with high and low PIs in each cancer. Patients of 13 cancers (including BLCA, BRCA, COAD, ESCA, HNSC, KIRC, KIRP, LIHC, LUAD, PAAD, PRAD, READ and UCSC) can be significantly separated into two groups (log-rank test, $P < 0.05$) based on 55 survival-related hypermethylated PDMCs (Figure 5C), and patients of eight cancers (including BRCA, COAD, HNSC, KIRC, KIRP, LUAD, LUSC and UCEC) can be significantly divided into two groups (log-rank test, $P < 0.05$) based on seven survival-related hypomethylated PDMCs (Supplementary Figure S17B). These survival-related hyper- and hypomethylated PDMCs can be prognostic signatures in many cancers. Additionally, we also perform survival analysis with the presence of patient age and tumor stage as confounders, and find that these confounders do not affect our conclusions (Supplementary Table S8).

csDMCs tend to be associated with cancer-specific functions

We obtain a set of hypermethylated csDMCs (from 176 in THCA to 4609 in UCEC) and hypomethylated csDMCs (from 123 in LUAD to 9331 in LIHC) (Supplementary Table S9) and investigate their associations with known cancer genes, cell-type-specific hypoMarks and cell-type-specific super-enhancers (Supplementary Materials). Both hyper- and hypomethylated csDMC-associated genes are significantly enriched in known cancer genes in most cancers (Supplementary Figure S18A and Table S9). For example, 138 of the 898 BRCA-specific hypermethylated genes are known breast cancer ones (FC = 2.03, P = 3.33e-16). Of the 205 PAAD-specific hypomethylated genes, 41 are known pancreatic cancer ones (FC = 2.72, P = 1.15e-9). Recently, Liu *et al.* [7] revealed a set of hypoMarks that are supposed to play roles in cell identity. Our hypermethylated csDMCs are significantly enriched in the cell-type-specific hypoMarks (Supplementary Figure S19 and Table S10). For instance, 64 of 295 PAAD-specific DMCs are overlapped with pancreas-specific hypoMarks (FC = 23.5, P = 7.29e-66). This confirms that the abnormal DNA methylation of csDMCs plays important roles in carcinogenesis.

Discussion and conclusion

Investigating DNA methylation patterns across multiple cancer types is essential for understanding the mechanisms of DNA methylation in tumorigenesis. In this work, we comprehensively investigate the common and specific DNA methylation patterns among DNA methylation profiles across 15 cancer types from TCGA. We also determine hyper- and hypomethylated PDMCs and csDMCs and explore their biological and clinical relevance with cancers. Extensive analysis indicates that these abnormal DNA methylation patterns are involved in tumor pathogenesis. For example, the TFs binding to hypermethylated PDMCs are likely relating to cancer hallmark pathways (Figure 3). While csDMC-associated genes are significantly relating to known cancer genes (Supplementary Figure S18A), we also report 55 hypermethylated PDMCs and seven hypomethylated PDMCs whose methylation levels are significantly correlated with patient survival in many cancers (Figure 5 and Supplementary Figure S17), which can be considered as prognostic signatures to provide valuable information for clinical treatment.

Recent studies revealed that DNA methylation of lincRNA promoter is closely correlated with lincRNA transcription and abnormal DNA methylation of lincRNA is associated with cancer status, subtypes or prognosis [75]. This indicates that abnormal DNA methylation of noncoding genes may serve as new biomarkers of cancer diagnosis. Many of the PDMCs are annotated in noncoding genes (Supplementary Table S11), and several have been proved to play important roles in cancer development. For example, eight hypermethylated PDMCs are annotated within TSS1500 of *MIR129-2*, which play as a tumor suppressor in colorectal cancer [76], lung cancer [77] and esophageal cancer [78]. Three hypomethylated PDMCs are annotated to the body region of noncoding genes *MEG3*, which suppress tumor growth in many cancers [79]. All of these

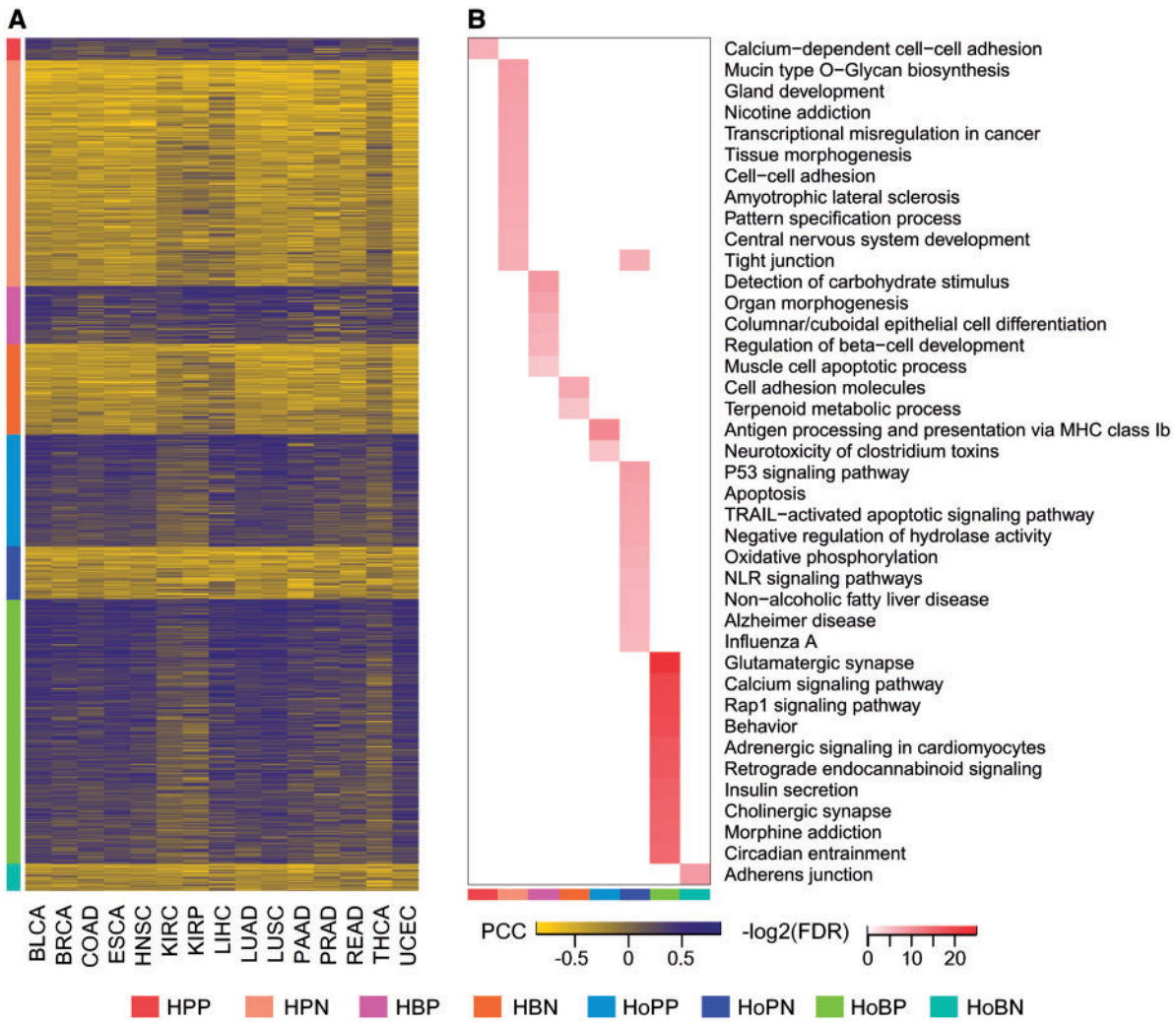

Figure 4. Pan-cancer-wide methylation-expression correlations. (**A**) The PCCs of the expression with methylation of genes in eight different groups. HPP = hypermethylated, promoter, positive; HPN = hypermethylated, promoter, negative; HBP = hypermethylated, body, positive; HBN = hypermethylated, body, negative; HoPP = hypomethylated, promoter, positive; HoPN = hypomethylated, promoter, negative; HoBP = hypomethylated, body, positive; HoBN = hypomethylated, body, negative. (**B**) Enriched pathways of eight different gene groups. For visualization, we chose the top 10 enriched pathway in each gene group. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

suggest that investigating the functions of PDMCs associated noncoding genes will be a valuable direction for future study.

Previous works have suggested that differentially methylated regions (DMRs) are more strongly correlated with gene expression [80]. However, DMRs are likely detected in CpG-rich regions, but seldom in CpG-deficient regions (e.g. the open sea regions). In this study, we aim to perform a comprehensive study of DNA methylation patterns across the whole genome. Moreover, HM450 data only contains about 480 000 CpGs (~2% of CpGs in the whole genome) [5]. Thus, we perform a probe-level study here. Additionally, many biological meaningful results have been discovered by CpG site/probe-specific DNA methylation studies. For example, Yao *et al.* [27] identified the target genes of enhancers based on a CpG probe-level study. Fleischer *et al.* [3] identified 18 CpGs associated with patient survival of breast cancer based on probe-specific DNA methylation. Moreover, we compute the distances of each neighboring hyper- or hypomethylated PDMCs, and find that many of the distances are >500 kb (Supplementary Figure S20). More specifically, 293 hyper- and 589 hypomethylated PDMCs are far from their nearest PDMCs at a distance of >500 kb, and these PDMCs may not be detected in a DMR-based study. We detect DMRs using *minfi* package [80] to validate this, and find that those DMRs only cover about 5.74% of all CpGs in the HM450 array (from 0.69% in THCA to 11.6% in UCEC), and only 3608 of 5450 hyper- and 2094 of 4433 hypomethylated PDMCs are covered in DMRs of at least eight cancer types. This indicates that many of our results cannot be discovered based on a DMR-based study.

Previous studies have demonstrated that usual processes, such as aging, would likely induce hypermethylation of Polycomb-bound sites [81]. To confirm whether the hypermethylation of the three enhancer probes (cg22620090, cg02829743 and cg02391713), which are at a target region of *EZH2*, is due to aging or tumorigenesis, we computed the PCCs between methylation levels of these three enhancer probes with patient ages across 15 cancer types (Supplementary Table S12). Most (at least 10 of 15) of the correlations are smaller than 0.1 ($P > 0.05$), indicating that aging is not the main reason of the hypermethylation of these three enhancer probes, and confirming that the hypermethylation should be owing to tumorigenesis.

Yang *et al.* [82] performed a pan-cancer study to identify the regulators of genome-wide DNA methylations, while little

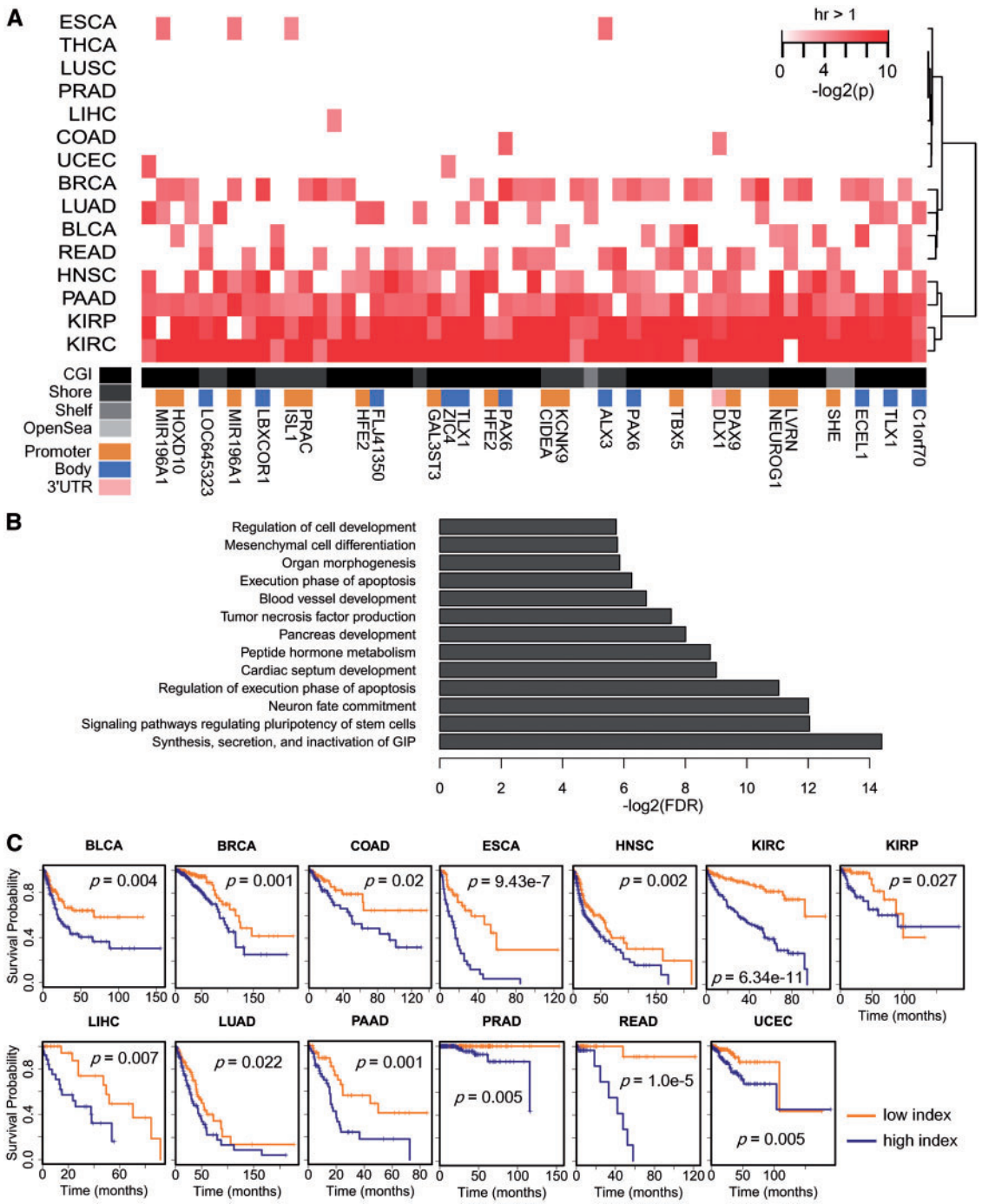

Figure 5. Fifty-five survival-related hypermethylated PDMCs. (**A**) The significance of these PDMCs correlating with patient survival time in 15 cancers. The P-value is computed by Wald test. 'hr' represents the hazard ratio. For convenience, we truncate the –log2(*p*) by 10. We label their relative positions (CGI, Shore, Shelf, Open sea) and associated genes with promoter, body or 3′UTR labels. (**B**) The pathway enrichment of the survival-related hypermethylated PDMC-associated genes. (**C**) Kaplan–Meier survival curves 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. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

concerns on the investigation of DNA methylation patterns in a pan-cancer view. They applied *t*-test to detect the differentially expressed epigenetic enzyme (EE) genes and a causal network model to identify the DNA methylation driver regulators from the differentially expressed EE genes. Interestingly, they identified that *EZH2* is one of the oncogenic EE genes and is correlated with both hyper- and hypomethylation. In our work, we find that *EZH2* is a key TF that associated with the regulation of enhancer regions we identified to the downstream tumor suppressor genes, *BVES* and *PRDM1* (Figure 2D). This may suggest that the oncogenic features of *EZH2* is owing to the loss of bindings, which induces hypermethylation of this region. This is an interesting problem worthy of further exploration.

We detected an enhancer region (containing cg22620090, cg02829743 and cg02391713) that hypermethylated in at least 11 cancer types (Figure 2A and Supplementary Figure S11), and this

region is exclusively bound by *EZH2* in normal cell lines (Supplementary Figure S9). Additionally, we identified two tumor suppressor genes, *BVES* and *PRDM1*, as the target genes of this region via associating these genes' expression with CpG methylations in this region (Figure 2C). We therefore conjectured that *EZH2* might bind to this region in normal cells to contribute to the repression of tumorigenesis by activating the target genes (Figure 2D). This seems contrary to the well-known repressive function of *EZH2*, observed as diverse as plants [83], human normal cells [84] and tumorigenesis cells [85]. However, recent studies have indicated that *EZH2* may also act as an activator for transcriptions [86]. In addition, H3K27me3, a well-known repressive marker that is catalyzed by *EZH2* [87], has also been identified to be associated with active transcription [88]. These results further confirm our conjecture and suggest the complexity of the functions of *EZH2*. A model system of *BMI1* and *CDKN2A/p16* [89] has been well established to model the 'Polycomb switching' on promoters, which may provide a clue for validating this 'switching' on enhancers. We need to note that this result is observed based on the computational study from large-scale data, which provides high potential interests for biologists to further validate.

In this work, we detect the PDMCs across 15 cancer types and investigate the correlations between methylation and expression of PDMC-associated genes, and the correlations between patient survival and methylation of PDMCs. We identify 55 hypermethylated and 7 hypomethylated PDMCs are significantly associated with patient survival times (Figure 5 and Supplementary Figure S17). Twenty-seven of the 55 hypermethylated PDMCs are associated with 24 genes and three of the seven hypomethylated PDMCs are associated with three genes. Interestingly, we find that 8 of the 27 genes (*C1orf70*, *GAL3ST3*, *ISL1*, *PAX6*, *PAX9*, *TLX1*, *ZIC4* and *TMEM132C*) show significantly correlations between methylation and gene expression (Supplementary Figure S21). All of these eight genes have been reported to be associated with cancers in previous works (Supplementary Table S7), and six of them have been reported as tumor suppressor genes or oncogenes. Interestingly, two tumor suppressor genes (*GAL3ST3* and *ISL1*) show hypermethylation and negatively correlations with their expression, and three oncogenes (*PAX6*, *PAX9* and *TLX1*) show hypermethylation and positively correlations with their expression, which indicate that the PDMCs in cancers tend to regulate the expressions of tumor suppressors or oncogenes to contribute to tumorigenesis and affect patient survival.

Key Points

- We identify differentially methylated CpG sites (DMCs) across 15 cancer types from TCGA and obtain 5450 hyper- and 4433 hypomethylated pan-cancer DMCs (PDMCs).
- We find an enhancer region that is hypermethylated in at least 11 cancer types. It regulates two tumor suppressor genes, *BVES* and *PRDM1*, negatively.
- Motifs enriched in hyper- or hypomethylated PDMCs are clustered into six clusters, which are associated with several well-known cancer hallmarks.
- We identify 55 hypermethylated and seven hypomethylated PDMCs that are significantly associated with patient survival.
- Known cancer genes and cell-type-specific super-enhancers are enriched in the cancer-specific DMCs.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgments

X.Y. would like to thank the support of National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, CAS during his visit.

Funding

This work was supported by the National Natural Science Foundation of China (No. 61379092, 61422309 to S.Z.; 61532014, 91530113, 61432010 to L.G.); the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) (XDB13040600), the Outstanding Young Scientist Program of CAS and the Key Laboratory of Random Complex Structures and Data Science at CAS to S.Z.; the Fundamental Research Funds for the Central Universities (No. BDZ021404 to L.G.).

References

1. Varley KE, Gertz J, Bowling KM, *et al*. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res* 2013;**23**:555–67.
2. Herman JG, Latif F, Weng Y, *et al*. Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. *Proc Natl Acad Sci USA* 1994;**91**:9700–4.
3. Fleischer T, Frigessi A, Johnson KC, *et al*. Genome-wide DNA methylation profiles in progression to in situ and invasive carcinoma of the breast with impact on gene transcription and prognosis. *Genome Biol* 2014;**15**:435.
4. Hinoue T, Weisenberger DJ, Lange CP, *et al*. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res* 2012;**22**:271–82.
5. Plongthongkum N, Diep DH, Zhang K. Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat Rev Genet* 2014;**15**:647–61.
6. Mendizabal I, Yi SV. Whole-genome bisulfite sequencing maps from multiple human tissues reveal novel CpG islands associated with tissue-specific regulation. *Hum Mol Genet* 2016;**25**:69–82.
7. Liu H, Liu X, Zhang S, *et al*. Systematic identification and annotation of human methylation marks based on bisulfite sequencing methylomes reveals distinct roles of cell type-specific hypomethylation in the regulation of cell identity genes. *Nucleic Acids Res* 2016;**44**:75–94.
8. Witte T, Plass C, Gerhauser C. Pan-cancer patterns of DNA methylation. *Genome Med* 2014;**6**:66.
9. Yang X, Shao X, Gao L, *et al*. Comparative DNA methylation analysis to decipher common and cell type-specific patterns among multiple cell types. *Brief Funct Genomics* 2016;doi: 10.1093/bfgp/elw013.
10. Yang X, Shao X, Gao L, *et al*. Systematic DNA methylation analysis of multiple cell lines reveals common and specific patterns within and across tissues of origin. *Hum Mol Genet* 2015;**24**:4374–84.
11. Gevaert O, Tibshirani R, Plevritis SK. Pancancer analysis of DNA methylation-driven genes using MethylMix. *Genome Biol* 2015;**16**:17.

12. Feng Q, Hawes SE, Stern JE, *et al*. DNA methylation in tumor and matched normal tissues from non-small cell lung cancer patients. *Cancer Epidemiol Biomarkers Prev* 2008;**17**:645–54.
13. Luo D, Huang H, Lu ML, *et al*. Abnormal expression of adhesion protein Bves is associated with gastric cancer progression and poor survival. *Pathol Oncol Res* 2012;**18**:491–7.
14. Williams CS, Zhang B, Smith JJ, *et al*. BVES regulates EMT in human corneal and colon cancer cells and is silenced via promoter methylation in human colorectal carcinoma. *J Clin Invest* 2011;**121**:4056–69.
15. Kang HB, Lee HR, Jee da J, *et al*. PRDM1, a tumor-suppressor gene, is induced by Genkwadaphnin in human colon cancer SW620 Cells. *J Cell Biochem* 2016;**117**:172–9.
16. Sandoval J, Heyn H, Moran S, *et al*. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* 2011;**6**:692–702.
17. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74.
18. Guintivano J, Aryee MJ, Kaminsky ZA. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics* 2013;**8**:290–302.
19. Jung N, Dai B, Gentles AJ, *et al*. An LSC epigenetic signature is largely mutation independent and implicates the HOXA cluster in AML pathogenesis. *Nat Commun* 2015;**6**:8489.
20. Futreal PA, Coin L, Marshall M, *et al*. A census of human cancer genes. *Nat Rev Cancer* 2004;**4**:177–83.
21. Abbott KL, Nyre ET, Abrahante J, *et al*. The candidate cancer gene database: a database of cancer driver genes from forward genetic screens in mice. *Nucleic Acids Res* 2015;**43**:D844–8.
22. Hnisz D, Abraham BJ, Lee TI, *et al*. Super-enhancers in the control of cell identity and disease. *Cell* 2013;**155**:934–47.
23. Smyth GK. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer, New York, 2005, 397–420.
24. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.
25. Heinz S, Benner C, Spann N, *et al*. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010;**38**:576–89.
26. Reimand J, Kull M, Peterson H, *et al*. g:Profiler–a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* 2007;**35**:W193–200.
27. Yao L, Shen H, Laird PW, *et al*. Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome Biol* 2015;**16**:105.
28. Ernst J, Kheradpour P, Mikkelsen TS, *et al*. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 2011;**473**:43–9.
29. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. Springer Science & Business Media, New York, 2000.
30. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012;**487**:330–7.
31. Goering W, Kloth M, Schulz WA. DNA methylation changes in prostate cancer. *Methods Mol Biol* 2012;**863**:47–66.
32. Moore LE, Pfeiffer RM, Poscablo C, *et al*. Genomic DNA hypomethylation as a biomarker for bladder cancer susceptibility in the spanish bladder cancer study: a case-control study. *Lancet Oncol* 2008;**9**:359–66.
33. Ruike Y, Imanaka Y, Sato F, *et al*. Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. *BMC Genomics* 2010;**11**:137.
34. Kamalakaran S, Varadan V, Giercksky Russnes HE, *et al*. DNA methylation patterns in luminal breast cancers differ from non-luminal subtypes and can identify relapse risk independent of other clinical variables. *Mol Oncol* 2011;**5**:77–92.
35. Fleischmann KK, Pagel P, Schmid I, *et al*. RNAi-mediated silencing of MLL-AF9 reveals leukemia-associated downstream targets and processes. *Mol Cancer* 2014;**13**:27.
36. Ossandon FJ, Villarroel C, Aguayo F, *et al*. In silico analysis of gastric carcinoma serial analysis of gene expression libraries reveals different profiles associated with ethnicity. *Mol Cancer* 2008;**7**:22.
37. Chen LF. Tumor suppressor function of RUNX3 in breast cancer. *J Cell Biochem* 2012;**113**:1470–7.
38. Bae SC, Choi JK. Tumor suppressor activity of RUNX3. *Oncogene* 2004;**23**:4336–40.
39. Chen F, Wang M, Bai J, *et al*. Role of RUNX3 in suppressing metastasis and angiogenesis of human prostate cancer. *PLoS One* 2014;**9**:e86917.
40. Creyghton MP, Cheng AW, Welstead GG, *et al*. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA* 2010;**107**:21931–6.
41. Roadmap Epigenomics C, Kundaje A, Meuleman W, *et al*. Integrative analysis of 111 reference human epigenomes. *Nature* 2015;**518**:317–30.
42. Dixon JR, Selvaraj S, Yue F, *et al*. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012;**485**:376–80.
43. Han P, Fu Y, Luo M, *et al*. BVES inhibition triggers epithelial-mesenchymal transition in human hepatocellular carcinoma. *Dig Dis Sci* 2014;**59**:992–1000.
44. Mandelbaum J, Bhagat G, Tang H, *et al*. BLIMP1 is a tumor suppressor gene frequently disrupted in activated B cell-like diffuse large B cell lymphoma. *Cancer Cell* 2010;**18**:568–79.
45. Rosenbloom KR, Armstrong J, Barber GP, *et al*. The UCSC genome browser database: 2015 update. *Nucleic Acids Res* 2015;**43**:D670–81.
46. Lee KH, Kwak YD, Kim DH, *et al*. Human zinc finger protein 161, a novel transcriptional activator of the dopamine transporter. *Biochem Biophys Res Commun* 2004;**313**:969–76.
47. Moreno-Smith M, Lee SJ, Lu C, *et al*. Biologic effects of dopamine on tumor vasculature in ovarian carcinoma. *Neoplasia* 2013;**15**:502–10.
48. Chakroborty D, Sarkar C, Yu H, *et al*. Dopamine stabilizes tumor blood vessels by up-regulating angiopoietin 1 expression in pericytes and kruppel-like factor-2 expression in tumor endothelial cells. *Proc Natl Acad Sci USA* 2011;**108**:20730–5.
49. Cao H, Yang ZX, Jiang GQ. Expression and clinical significance of activating transcription factor 3 in human breast cancer. *Iran J Basic Med Sci* 2013;**16**:1151–4.
50. Wolford CC, McConoughey SJ, Jalgaonkar SP, *et al*. Transcription factor ATF3 links host adaptive response to breast cancer metastasis. *J Clin Invest* 2013;**123**:2893–906.
51. Song X, Lu F, Liu RY, *et al*. Association between the ATF3 gene and non-small cell lung cancer. *Thorac Cancer* 2012;**3**:217–23.
52. Mahony S, Benos PV. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res* 2007;**35**:W253–8.

53. Yu G, Li F, Qin Y, *et al*. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 2010;**26**:976–8.
54. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011;**144**:646–74.
55. Tachibana K, Yamasaki D, Ishimoto K, *et al*. The role of PPARs in cancer. *PPAR Res* 2008;**2008**:102737.
56. Amati B, Land H. Myc-Max-Mad: a transcription factor network controlling cell cycle progression, differentiation and death. *Curr Opin Genet Dev* 1994;**4**:102–8.
57. Han H, Shim H, Shin D, *et al*. TRRUST: a reference database of human transcriptional regulatory interactions. *Sci Rep* 2015;**5**:11432.
58. Karan D, Chen SJ, Johansson SL, *et al*. Dysregulated expression of MIC-1/PDF in human prostate tumor cells. *Biochem Biophys Res Commun* 2003;**305**:598–604.
59. Scian MJ, Stagliano KE, Deb D, *et al*. Tumor-derived p53 mutants induce oncogenesis by transactivating growth-promoting genes. *Oncogene* 2004;**23**:4430–43.
60. Pils D, Bachmayr-Heyda A, Auer K, *et al*. Cyclin E1 (CCNE1) as independent positive prognostic factor in advanced stage serous ovarian cancer patients— a study of the OVCAD consortium. *Eur J Cancer* 2014;**50**:99–110.
61. Stegh AH. Targeting the p53 signaling pathway in cancer therapy— the promises, challenges and perils. *Expert Opin Ther Targets* 2012;**16**:67–83.
62. Wang S, El-Deiry WS. TRAIL and apoptosis induction by TNF-family death receptors. *Oncogene* 2003;**22**:8628–33.
63. Largaespada DA. A bad rap: Rap1 signaling and oncogenesis. *Cancer Cell* 2003;**4**:3–4.
64. Young MD, Wakefield MJ, Smyth GK, *et al*. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* 2010;**11**:R14.
65. Geeleher P, Hartnett L, Egan LJ, *et al*. Gene-set analysis is severely biased when applied to genome-wide methylation data. *Bioinformatics* 2013;**29**:1851–7.
66. Lee DW, Han SW, Cha Y, *et al*. Different prognostic effect of CpG island methylation according to sex in colorectal cancer patients treated with adjuvant FOLFOX. *Clin Epigenetics* 2015;**7**:63.
67. Kim JH, Cho NY, Bae JM, *et al*. Nuclear maspin expression correlates with the CpG island methylator phenotype and tumor aggressiveness in colorectal cancer. *Int J Clin Exp Pathol* 2015;**8**:1920–8.
68. Yang Q, Shao Y, Shi J, *et al*. Concomitant PIK3CA amplification and RASSF1A or PAX6 hypermethylation predict worse survival in gastric cancer. *Clin Biochem* 2014;**47**:111–16.
69. Yu J, Ma X, Cheung KF, *et al*. Epigenetic inactivation of T-box transcription factor 5, a novel tumor suppressor gene, is associated with colon cancer. *Oncogene* 2010;**29**:6464–74.
70. Suzuki H, Takatsuka S, Akashi H, *et al*. Genome-wide profiling of chromatin signatures reveals epigenetic regulation of MicroRNA genes in colorectal cancer. *Cancer Res* 2011;**71**:5646–58.
71. Jadhav RR, Ye Z, Huang RL, *et al*. Genome-wide DNA methylation analysis reveals estrogen-mediated epigenetic repression of metallothionein-1 gene cluster in breast cancer. *Clin Epigenetics* 2015;**7**:13.
72. Cogill SB, Wang L. Co-expression network analysis of human lncRNAs and cancer genes. *Cancer Inform* 2014;**13**:49–59.
73. Martin J, St-Pierre MV, Dufour JF. Hit proteins, mitochondria and cancer. *Biochim Biophys Acta* 2011;**1807**:626–32.
74. Kampen KR. Membrane proteins: the key players of a cancer cell. *J Membr Biol* 2011;**242**:69–74.
75. Zhi H, Ning S, Li X, *et al*. A novel reannotation strategy for dissecting DNA methylation patterns of human long intergenic non-coding RNAs in cancers. *Nucleic Acids Res* 2014;**42**:8258–70.
76. Karaayvaz M, Zhai H, Ju J. miR-129 promotes apoptosis and enhances chemosensitivity to 5-fluorouracil in colorectal cancer. *Cell Death Dis* 2013;**4**:e659.
77. Xiao Y, Li X, Wang H, *et al*. Epigenetic regulation of miR-129-2 and its effects on the proliferation and invasion in lung cancer cells. *J Cell Mol Med* 2015;**19**:2172–80.
78. Kang M, Li Y, Liu W, *et al*. miR-129-2 suppresses proliferation and migration of esophageal carcinoma cells through down-regulation of SOX4 expression. *Int J Mol Med* 2013;**32**:51–8.
79. Zhou Y, Zhang X, Klibanski A. MEG3 noncoding RNA: a tumor suppressor. *J Mol Endocrinol* 2012;**48**:R45–53.
80. Aryee MJ, Jaffe AE, Corrada-Bravo H, *et al*. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. *Bioinformatics* 2014;**30**:1363–9.
81. Jung M, Pfeifer GP, Aging and DNA methylation. *BMC Biol* 2015;**13**:7.
82. Yang Z, Jones A, Widschwendter M, *et al*. An integrative pan-cancer-wide analysis of epigenetic enzymes reveals universal patterns of epigenomic deregulation in cancer. *Genome Biol* 2015;**16**:140.
83. Angel A, Song J, Dean C, *et al*. A polycomb-based switch underlying quantitative epigenetic memory. *Nature* 2011;**476**:105–8.
84. Xu J, Shao Z, Li D, *et al*. Developmental control of polycomb subunit composition by GATA factors mediates a switch to non-canonical functions. *Mol Cell* 2015;**57**:304–16.
85. Lund K, Adams PD, Copland M. EZH2 in normal and malignant hematopoiesis. *Leukemia* 2014;**28**:44–9.
86. Tan JZ, Yan Y, Wang XX, *et al*. EZH2: biology, disease, and structure-based drug discovery. *Acta Pharmacol Sin* 2014;**35**:161–74.
87. Shema E, Jones D, Shoresh N, *et al*. Single-molecule decoding of combinatorially modified nucleosomes. *Science* 2016;**352**:717–21.
88. Young MD, Willson TA, Wakefield MJ, *et al*. ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Res* 2011;**39**:7415–27.
89. Park IK, Morrison SJ, Clarke MF. Bmi1, stem cells, and senescence regulation. *J Clin Invest* 2004;**113**:175–9.