Molecular Oncology

FEBS PRESS science publishing by scientists

# Identification and validation of novel DNA methylation markers for early diagnosis of lung adenocarcinoma

Miao Li[1], Chen Zhang[1], Lijun Zhou[1], Siyu Li[1], Yuan Jie Cao[2], Longlong Wang[1,3], Rong Xiang[1], Yi Shi[1,3] and Yongjun Piao[1,3]

1 School of Medicine, Nankai University, Tianjin, China

2 Department of Radiation and Oncology, National Clinical Research Center for Cancer and Tianjin Key Laboratory of Cancer Prevention and Therapy, Tianjin Medical University Cancer Institute and Hospital, Tianjin, China

3 Tianjin Key Laboratory of Human Development and Reproductive Regulation, Nankai University Affiliated Hospital of Obstetrics and Gynecology, Tianjin, China

**Keywords**
DNA methylation; early diagnosis; feature selection; lung adenocarcinoma

**Correspondence**
Y. Piao and Y. Shi, School of Medicine, Nankai University, Tianjin 300071, China
E-mails: ypiao@nankai.edu.cn (YP) and yishi@nankai.edu.cn (YS)

(Received 4 March 2020, revised 7 June 2020, accepted 16 July 2020, available online 27 August 2020)

doi:10.1002/1878-0261.12767

Lung cancer has the highest mortality of all cancers worldwide. Epigenetic alterations have emerged as potential biomarkers for early diagnosis of various cancer tissue types. To identify methylation markers for early diagnosis of lung adenocarcinoma, we aimed to integrate genome-wide DNA methylation and gene expression data from The Cancer Genome Atlas. To this end, we first examined the global DNA methylation pattern of lung adenocarcinoma and investigated the relationship between DNA methylation subtypes and clinical features. We then extracted differentially methylated and expressed genes, and adopted feature selection techniques to determine the final methylation markers. The performance of the markers in predicting lung adenocarcinoma was evaluated on three independent datasets from Gene Expression Omnibus. Protein levels of marker genes were validated by immunohistochemistry, and their biological function was further verified *in vivo*. We identified three novel methylation markers in lung adenocarcinoma including cg08032924, cg14823851, and cg19161124, mapping to *CMTM2*, *TBX4*, and *DPP6*, respectively. Validating these results on three independent datasets indicated that the three markers can achieve extremely high sensitivity and specificity in distinguishing lung adenocarcinoma from normal samples. Immunohistochemistry quantification results confirmed that markers are weakly expressed in human lung adenocarcinoma, and *CMTM2* decreased tumor growth of mouse Lewis lung carcinoma *in vivo*. Overall, our study identified three novel methylation markers in lung adenocarcinoma which may contribute toward an improved diagnosis potentially leading to a better outcome for patients with lung adenocarcinoma.

**Abbreviations**
1st exon, the first exon; AUC, area under the curve; BMP, bone morphogenetic protein; CGI, CpG islands; CIMP, CpG island methylation phenotype; DEG, differentially expressed gene; DMC, differentially methylated cytosine; GEO, gene expression omnibus; GR, gain ratio; IG, information gain; IHC, immunohistochemistry; LUAD, lung adenocarcinoma; NSCLC, non-small-cell lung cancer; RF, reliefF; ROC, receiver operating characteristic; SU, symmetrical uncertainty; TCGA, the cancer genome atlas.

## 1. Introduction

Lung cancer is the leading cause of death with cancer worldwide [1]. An estimated 72 000 deaths in men and an estimated 63 220 deaths in women from lung cancer occurred in the United States alone in 2020 [2]. As the major histological type, non-small-cell lung cancer (NSCLC) accounts for ~ 80% of all lung cancer cases, in which lung adenocarcinoma (LUAD), arising from the mucus-secreting glandular cells, accounts for ~ 50% [3]. The average 5-year survival rate in the United States for the patients diagnosed with lung cancer during 2009 through 2015 was as low as 19%. It is noticeable that the survival rates varied most among different stages of lung cancer. The 5-year relative survival rate is only 5% for patients diagnosed with metastatic disease, which is far less than the rate for patients diagnosed with localized stage disease (57%) [2]. Obviously, early detection contributes to favorable prognosis, and thus, early screening and diagnosis of cancer is of great significance.

Cancer screening tests have been used to detect different types of cancers at an early stage. The National Lung Screening Trial (NLST) showed that low-dose helical computed tomography (CT) screening can reduce lung cancer mortality [4,5]. However, not all of the cancers detected by screening with low-dose CT will be found early. Furthermore, although low-dose CT is a diagnostic method with high sensitivity, it often detects things that turn out not to be cancer [6]. Further follow-up or invasive tests are required after screening for accurate diagnosis. As the advance of high-throughput technologies, such as next-generation sequencing, epigenetic alterations [7–9] of oncogenes or tumor suppressor genes have been investigated and emerged as the potential biomarkers for early diagnosis of cancers.

DNA methylation is a chemical modification of DNA by which methyl groups are added to the cytosines [10,11]. Hypermethylation of tumor suppressor genes is a common event in various tumors, suggesting that DNA methylation alterations could be a new strategy for cancer diagnosis [12]. Compared with protein and genetic markers, DNA methylation signatures have a number of advantages. Methylation markers are relatively sensitive and stable than protein markers [12]. DNA methylation alterations often occur in the early stage of the cancer [13]. Moreover, methylation signatures can be detected in both cancer tissue and circulating tumor DNA which can be obtained in a minimally invasive manner [14]. In recent years, several candidate methylation markers have been studied in various cancer types [15–17], but none has been used in clinical practice yet. Integrative analysis of genome-wide DNA methylation and gene expression has become an alternative method for systematically understanding the role of methylation variation in cancers with the potential of discovering new epigenetic markers that are more sensitive and robust.

In this study, we performed an integrative analysis of genome-wide DNA methylation and gene expression data to identify DNA methylation markers for early diagnosis of LUAD. For this purpose, we used two public databases: The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO). Using a machine learning approach, we finally identified three LUAD methylation markers including cg08032924, cg14823851, and cg19161124, mapped to *CMTM2*, *TBX4*, and *DPP6*, respectively. A logistic regression model based on the combination of these markers can accurately distinguish LUAD from normal samples on independent validation sets. The protein expression patterns of the markers were further validated by immunohistochemistry, and the suppression of tumor growth of *CMTM2* was confirmed in the mouse model.

## 2. Methods

### 2.1. DNA methylation

Illumina HumanMethylation450K array data of 415 LUAD and 31 associated normal tissues were downloaded from UCSC Xena (cohort: GDC TCGA Lung Adenocarcinoma) [18]. CpGs were annotated using human reference genome version 19 using IlluminaHumanMethylation450kanno.ilmn12.hg19 R package [19], and CpGs contain SNPs were removed from the analysis. Among the original data, the methylation profiles of paired adjacent normal tissues were available for 29 LUAD samples. Thus, these 29 primary tumors and matched adjacent normal samples were selected for differential methylation analysis. CpGs have missing values in less than 20% of the samples were imputed using mean methylation levels while those with more than 20% missing values were removed for differential analysis. The differentially methylated cytosines (DMCs) were reported with false discovery rate (FDR) < 0.05 (the Wilcoxon rank-sum test) and methylation difference > 0.2. In addition, three independent validation datasets including GSE114989, GSE83842, and GSE85845 were obtained from the GEO [20]. GSE114989 [21] included 27

primary tumors and 7 matched normal tissues from 7 LUAD patients, GSE83842 [22] contained 12 cases with paired tumor and normal tissue, and GSE85845 [23] included 8 LUAD and adjacent nontumor tissues.

### 2.2. Gene expression

The HTSeq counts of RNA-seq data for LUAD including 524 tumors and 59 normal samples were obtained from UCSC Xena, and the log-transformed counts were converted into raw counts. Of the data, 18 primary tumors and matched adjacent normal samples were selected for differential expression analysis. The raw counts were normalized using the trimmed mean of *M* values (TMM) method, and EDGER [24] was used to perform the differential analysis. Differentially expressed genes (DEGs) were determined with adjusted *P*-value < 0.05 and the log fold change > 1.5.

### 2.3. Clinical characteristics

The well-preprocessed clinical information of the LUAD patients was obtained from [25] including basic characteristics such as sex, age at diagnosis, tumor stage, smoking status, and mutation status of genes such as *STK11*, *KRAS*, *KEAP1*, and *EGFR*. The survival data of the patients were obtained from UCSC Xena, and the Kaplan–Meier analysis with log-rank test was used to compare overall survival across different groups.

### 2.4. Unsupervised clustering analysis

*K*-means clustering algorithm with the Euclidean distance was used to determine the methylation subtypes of LUAD. Of the 54 429 promoter CpGs, 35 414 CpGs that were methylated ($\beta > 0.05$) in 32 normal tissues were removed, 18 859 CpGs showing low variations ($\sigma < 0.2$) in 153 tumor samples were removed, and finally 156 most variable CpGs were retained for clustering analysis.

### 2.5. Statistical analysis

The associations of methylation subtypes with clinical characteristics including sex, age at diagnosis, smoking history, tumor stage, smoking history, *STK11* mutation, *KEAP1* mutation, *KRAS* mutation, and *EGFR* mutation were examined using Fisher's exact test. The Kruskal–Wallis test was used to assess the statistical significance of differences in mean methylation levels among clusters. Pearson's correlation analysis was performed to assess the relationship between methylation status of CpGs and expression levels of genes.

### 2.6. Functional annotation

Gene ontology analysis was performed using the DAVID functional annotation tool [26], and significantly enriched (FDR < 0.05) biological processes and molecular functions were reported.

### 2.7. Marker identification by feature selection

Information gain [27], gain ratio [28], symmetrical uncertainty [29], and reliefF [30] in WEKA software package [31] were used for initial screening of the methylation markers. Information gain, gain ratio, and symmetrical uncertainty were all entropy-based impurity measures. ReliefF considers differences in nearest neighbors to obtain the feature weights. All feature selection methods generate a score for each CpG that can be used to rank features. Top 15 scoring CpGs by each method were recorded. Default parameters were used for all methods except the number of selected attributes.

### 2.8. Immunohistochemical staining

Tissue microarrays (HLugA060PG02 and HLugA150CS03) were purchased from Shanghai Outdo Biotech Co., Ltd. (Shanghai, China), containing 105 human LUAD and paired normal adjacent lung tissues. The paraffin sections were stained with anti-*TBX4* antibody (#sc-515196; Santa Cruz Biotechnology, Shanghai, China) at a dilution of 1/10, anti-*CMTM2* antibody (#PA5-50208; Thermo Fisher Scientific, Shanghai, China) at a dilution of 1/150, and anti-*DPP6* antibody (#sc-365147; Santa Cruz Biotechnology) at a dilution of 1/50. For microwave antigen retrieval, Tris-EDTA Buffer (pH 8.0) was employed, and multiple antigen retrievals were used if necessary. The *H*-score was used for quantifying the protein levels in human LUAD and paired adjacent normal lung tissues. *H*-score, ranging from 0 to 300, is the sum over product, which is calculated by multiplying the percentage of positive cells at each intensity and its staining intensity (weak, moderate, and strong were scored as 1, 2, and 3 based on color density).

### 2.9. Establishment of stable cell lines

To construct pLV-EF1α-Cmtm2-IRES-Bsd, DNA sequences encoding murine Cmtm2 were amplified from cDNA extracted from murine testis, then were cloned

into the plasmid pLV-EF1α-MCS-IRES-Bsd (Biosettia, San Diego, CA, USA). Next, the lentiviruses carrying pLV-EF1α-Cmtm2-IRES-Bsd or pLV-EF1α-MCS-IRES-Bsd were packaged combined with commercial transfection agents, Lipofectamine 2000 (#11668027; Thermo Fisher Scientific). Mouse Lewis lung carcinoma (LLC) cells were incubated with the lentivirus-containing supernatant with the presence of 8 $\mu g \cdot mL^{-1}$ polybrene for 48 h and followed by a selection with 10 $\mu g \cdot mL^{-1}$ blasticidin for one week to establish stable cell lines. The primers for amplification are mCmtm2 Forward 5′-TCAACGCGTGCCACCATGGCAGCACCGATAAAGTTTCC-3′ and mCmtm2 Reverse 5′-TCAGCTAGCTTACCACTTCCTTAACCTA-3′.

### 2.10. Real-time quantitative PCR

Total RNA was extracted using TRIzol reagent (#15596026; Thermo Fisher Scientific) and then was under reverse transcription via the TransScript First-Strand cDNA Synthesis SuperMix Kit (TransGen Biotech, Beijing, China). Quantitative PCR was performed on Roche real-time PCR detection system using the following primers: Q-Cmtm2 Forward 5′-CCCAAAAAGGGGGCTTCGAC-3′, Q-Cmtm2 Reverse 5′-ACCGGATGTGGGAGCATTGT-3′.

### 2.11. Mouse model

Seven-week-old C57BL/6J mice (Vital River Laboratory Animal Technology Co. Ltd) were used and maintained in a specific pathogen-free facility. Luciferase-expressing Lewis lung carcinoma cells ($1.5 \times 10^5$) were injected subcutaneously into the right flank of C57BL/6J mice, and then, the volume of tumors was monitored every 3 days. And tumor weights were measured when the mice were sacrificed 4 weeks postimplant.

### 2.12. Bioluminescence imaging

For live imaging, the C57BL/6J mice were given intraperitoneal injections of the reporter substrate (15 $mg \cdot mL^{-1}$ stock in PBS, 100 $\mu g \cdot g^{-1}$ mouse) 10 min before imaging and then were transferred to the imaging chamber for imaging after anesthesia. Images were analyzed using Living Image software, and the fluorescence intensity was quantified.

### 2.13. Ethical approval

All the experiments involving mice were conducted according to the guidelines established by the Nankai University Animal Care and Use Committee (NUACUC) by skilled experimenters under an approved protocol, which was in accordance with the principles and procedures outlined in the NIH Guide for the Care and Use of Laboratory Animals.

## 3. Results

### 3.1. Global DNA methylation pattern in LUAD

To examine the global DNA methylation patterns in LUAD patients, the mean β value was calculated for each CpG dinucleotide across 415 tumors, and the distribution of methylation levels was examined in CpG islands (CGIs), shores, and shelves (Fig. 1A). A bimodal distribution was observed for all CpGs while a large hypomethylation (the peak in the left) was found for CpGs located in CGIs. In addition, the CpGs in both north and south shores had variable methylation levels (bimodal distribution) and the CpGs in north and south shelves had large hypermethylations (the peak in the right), indicating the CpGs within or near the CGIs tend to have low methylation levels, which is consistent with previous findings [32,33]. We then performed an unsupervised clustering analysis of 156 CpGs that varied most across 153 well-annotated LUAD samples (Fig. 1B). The DNA methylation profile of tumors was clustered into three distinct subtypes, which denoted C1 ($n = 62$), C2 ($n = 54$), and C3 ($n = 37$), and the mean beta values indicated a significant difference (Kruskal–Wallis $P < 2.2e\text{-}16$) among clusters (Fig. 1C). Then, we investigated the association between the clusters and clinical characteristics, and sex, tumor stage, smoking history, *STK11* mutation, and *KEAP1* mutation were significantly ($P$-value < 0.05) associated with clusters (Table 1). Kaplan–Meier survival analysis was performed to estimate overall survival of each cluster, and there were

**Fig. 1.** Genome-wide DNA methylation patterns in LUAD. (A) The distribution of mean methylation levels of CpGs across 415 LUAD patients in whole-genome CpG islands, north shores, south shores, north shelves, and south shelves. (B) Consensus clustering of 156 CpGs that varied most across 153 well-annotated LUAD samples. Samples are presented in columns, and the CpGs are presented in rows. The methylation profile was clustered into three groups denoted as C1 ($n = 62$), C2 ($n = 54$), and C3 ($n = 37$). (C) The distributions of methylation levels in three clusters. Kruskal–Wallis test. (D) The distribution of hypermethylated and hypomethylated CpGs in different genomic regions including CpG island, shore, shelf, TSS1500, TSS200, 5′UTR, 1stExon, body, and 3′UTR.

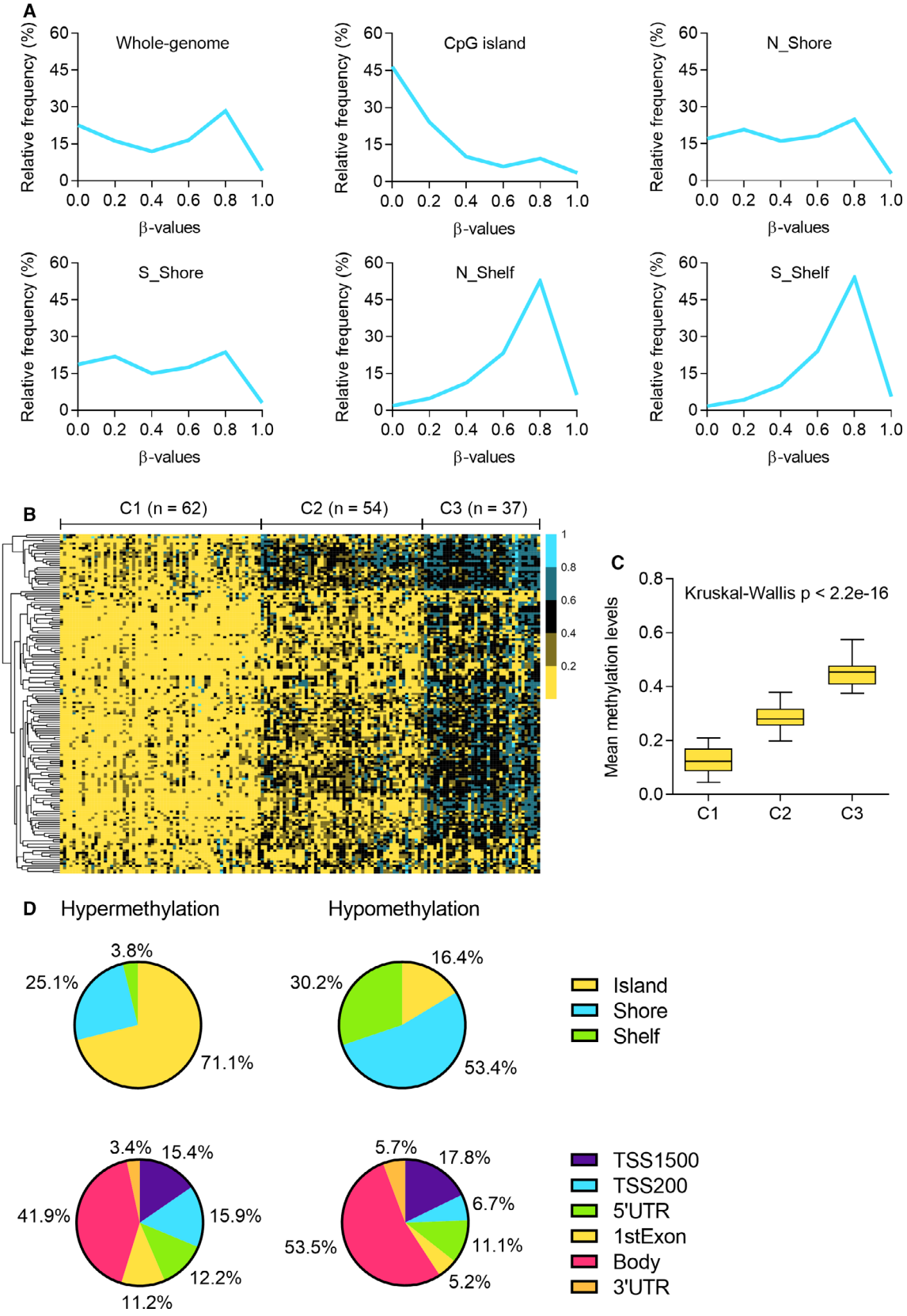
A
Whole-genome
CpG island
N_Shore
S_Shore
N_Shelf
S_Shelf
Relative frequency (%)
β-values
B
C1 (n = 62)
C2 (n = 54)
C3 (n = 37)
C
Kruskal-Wallis p < 2.2e-16
Mean methylation levels
C1
C2
C3
D
Hypermethylation
Hypomethylation
3.8%
25.1%
71.1%
16.4%
30.2%
53.4%
Island
Shore
Shelf
3.4%
15.4%
41.9%
15.9%
12.2%
11.2%
5.7%
17.8%
6.7%
53.5%
11.1%
5.2%
TSS1500
TSS200
5'UTR
1stExon
Body
3'UTR

**Table 1.** Clinical characteristics.

| Characteristics | Classes | C1 | C2 | C3 | *P*-value |
|---|---|---|---|---|---|
| Sex | Female | 34 | 37 | 15 | 0.030 |
| | Male | 28 | 17 | 22 | |
| Age at diagnosis[a] | ≥ 66 | 34 | 27 | 18 | 0.820 |
| | <66 | 28 | 27 | 19 | |
| Stage | Low (Stage I, Stage II) | 44 | 47 | 25 | 0.044 |
| | High (Stage III, Stage IV) | 18 | 7 | 12 | |
| Smoking history | Current or past smoker | 59 | 40 | 34 | 0.003 |
| | Lifelong nonsmoker | 3 | 14 | 3 | |
| *STK11* mutation | Mutant | 18 | 7 | 4 | 0.038 |
| | WT | 44 | 47 | 33 | |
| *KRAS* mutation | Mutant | 22 | 16 | 14 | 0.691 |
| | WT | 40 | 38 | 23 | |
| *KEAP1* mutation | Mutant | 18 | 5 | 4 | 0.014 |
| | WT | 44 | 49 | 33 | |
| *EGFR* mutation | Mutant | 10 | 7 | 4 | 0.832 |
| | WT | 52 | 47 | 33 | |

[a]Average age of the patients is 66.

no significant differences among different clusters (*P*-value = 0.46). Overall, hypermethylation of CGI was observed in LUAD patients as with in other cancer types, and the samples can clearly be divided into three methylation groups. However, high methylation was unlikely to be associated with poor survival.

Next, we analyzed the methylation differences in 29 LUAD and 29 matched normal samples. A total of 11 266 DMCs mapped to 3119 genes were detected, including 7415 hypermethylation (1687 genes) and 3851 hypomethylation (1432 genes) in tumors. We then investigated the distribution of hypermethylated and hypomethylated CpGs and genes in various genomic regions (Fig. 1D). Among hypermethylated CpGs, 71.1% were located in CGIs, 25.1% were in shores, and 3.8% were in shelves. However, 53.4% of hypermethylated CpGs were located in shores, 30% were in shelves, and only 16.4% were in CGIs. The variation of distribution between hypermethylated CpGs and hypomethylated ones was relatively smaller in gene-context regions than in CGI-based regions. Of hypermethylated (hypomethylated) CpGs, 15.4% (17.8%), 15.9% (6.7%), 12.2% (11.1%), 11.2% (5.2%), 41.9% (53.5%), and 3.4% (5.7%) were located in 1500-bp upstream of transcription start site (TSS1500), 200-bp upstream of TSS (TSS200), 5′ untranslated region (5′UTR), the first exon (1stExon), gene body, and 3′ untranslated region (3′UTR), respectively. It is obvious that the number of hypermethylated sites was higher in the regions near TSS. By differential methylation analysis, we identified 11 266 sites showing significant DNA methylation changes in tumors, and those DMCs were further used to be correlated with gene expression in downstream analysis to filter the DMCs that do not contribute to transcriptional regulation of genes.

### 3.2. Identification of relevant DNA methylation changes associated with mRNA expression

We performed an integrated analysis of DNA methylation and gene expression to identify potentially relevant DNA methylation alterations in LUAD. Of the 29 matched tumor and normal samples for differential methylation analysis, 18 pairs that have expression profiles were used for differential expression analysis. A total of 2622 DEGs were detected including 1500 upregulated genes and 1086 downregulated genes. Of these genes, approximately one fifth of them showed significant methylation changes between LUAD and normal samples, including 134 (487 CpGs) hypermethylated and upregulated genes, 147 (383 CpGs) hypermethylated and downregulated genes, 128 (211 CpGs) hypomethylated and upregulated genes, and 82 (160 CpGs) hypomethylated and downregulated genes (Fig. 2A). Gene Ontology (GO) analysis was then performed to examine the biological functions of the 147 hypermethylated and downregulated genes (Fig. 2B). In biological processes, negative regulation of cell growth, stem cell proliferation, positive regulation of *BMP* signaling pathway, and *Wnt* signaling pathway were significantly enriched. In terms of molecular function, the genes were related to transforming growth factor beta binding, transcriptional activator activity, sequence-specific binding, RNA polymerase II regulatory region sequence-specific DNA binding, and calcium ion binding. Cancer-related pathways such as *Wnt* signaling pathway were enriched in hypermethylated and downregulated groups.

Next, we performed a correlation analysis to assess the relationship between DNA methylation and gene expression. Pearson's correlation coefficients were calculated between 281 938 CpGs and corresponding genes (Fig. 2C). Using a coefficient cutoff of 0.3, 595 genes (19%) were positively correlated with methylation while 2409 genes (75%) were negatively correlated with methylation. Similar patterns were observed when considering correlations in DEGs, differentially methylated genes, and differentially expressed and methylated genes. Note that multiple CpGs can be associated with a gene, and the methylation of those CpGs can be both positively and negatively correlated with gene expression. Then, we examined the distribution of CpGs that are significantly correlated with

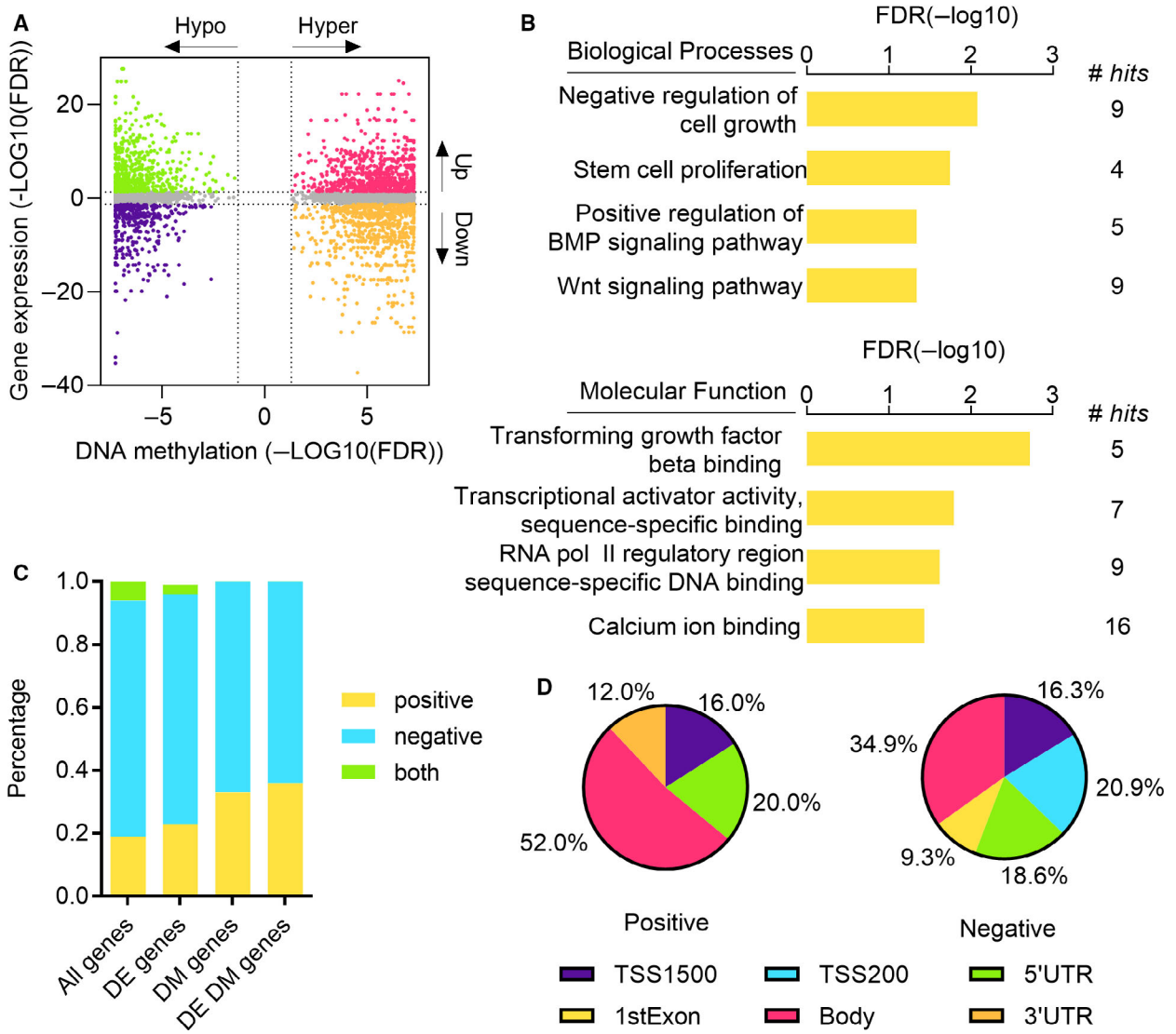


**Fig. 2.** Joint analysis of DNA methylation and mRNA expression. (A) Starburst plot integrating DNA methylation changes and gene expression changes ($n = 4297$). The genes are divided into four groups that are hypermethylated and upregulated (pink); hypermethylated and downregulated (orange); hypomethylated and upregulated (green); hypomethylated and downregulated in LUAD (blue). (B) GO analysis for hypermethylated and downregulated genes. (C) Percentage of positive/negative correlation between DNA methylation and gene expression. Pearson's correlation coefficient was calculated for all genes, DEGs, differentially methylated genes, and differentially expressed and methylated genes. (D) The distribution of positive and negative correlations in different genomic areas.

genes in different genomic regions (Fig. 2D). Of the CpGs positively affected gene expression, approximately half of them were located in gene body, 20% were in 5′UTR, 16% were in TSS1500, and 12% were in 3′UTR. Of the CpGs negatively regulated gene expression, 34.9% were located in gene body, 20.9% were in TSS200, 18.6% were in 5′UTR, 16.3% were in TSS1500, and 9.3% were in 1stExon. Interestingly, all CpGs in TSS200 and 1stExon were negatively correlated with gene expression. These results reveal that the CpGs located near TSS tend to negatively regulate expression of genes while the CpGs in gene body tend to positively regulate gene expression.

### 3.3. Identification and validation of methylation signatures in LUAD

To identify methylation markers for LUAD diagnosis, the candidate markers were further narrowed down to hypermethylated and downregulated CpGs since the

repression of tumor suppressor genes by the promoter hypermethylation is one of the most frequently observed epigenetic alterations in cancers. Of the 383 hypermethylated and downregulated CpGs, we selected 138 CpGs in promoter regions as a candidate functionally relevant group, and machine learning techniques were adopted to determine the final methylation markers in LUAD (Fig. 3A). We used four different feature selection approaches that were information gain (IG), gain ratio (GR), symmetrical uncertainty (SU), and reliefF (RF) to screen the markers. We extracted top 15 ranked CpGs found by each method and took the intersections of those CpGs as the final methylation markers (Table 2). Finally, we identified three methylation markers that were cg08032924, cg14823851, and cg19161124, mapped to *CMTM2*, *TBX4*, and *DPP6*, respectively (Fig. 3B). A logistic regression model was then built with these markers on TCGA LUAD samples, and the model was validated on three independent datasets from GEO. The areas under the receiver operating characteristic curve (AUCs) were 0.923, 1, and 0.905 for GSE114989, GSE83842, and GSE85845, respectively, indicating that the three markers can accurately classify LUAD samples from controls (Fig. 3C). To examine whether these markers were suitable for early detection of LUAD, we compared the methylation levels of TCGA LUAD patients in different tumor stages. All the markers were found to be significantly (Mann–Whitney $P$-value < 0.0001) hypermethylated in stage I tumors compared to normal samples (Fig. 3D). In addition, the patients were divided into high methylation and low methylation groups based on the average methylation levels of the markers, and the Kaplan–Meier analysis was conducted to investigate the association between the methylation status of the markers and the overall survival of patients (Fig. 3E). For cg08032924 and cg14823851, patients with low methylation levels had significantly better survival than those with high methylation levels ($P = 0.0367$ and $P = 0.0917$). For cg19161124, the overall survival between the two groups was not statistically significant ($P = 0.5953$) but the low methylation group still had better survival than the high methylation group. These results suggest that the identified markers can accurately predict LUAD and also work well on early-stage patients.

### 3.4. *CMTM2* and *TBX4* are weakly expressed in human LUAD

To further confirm the correlation between the three newly identified hypermethylated genes and LUAD progression, immunohistochemistry (IHC) was performed to evaluate the protein expression of *CMTM2*, *TBX4,* and *DPP6* in tissue arrays containing 105 human LUAD specimens and paired adjacent normal tissues. Quantitative analysis based on intact and paired specimens via *H*-score revealed that the expression of *CMTM2* and *TBX4* was obviously lower in LUAD when compared to the paired adjacent normal lung tissues, implying the potential gene silencing of *CMTM2* and *TBX4* caused by DNA methylation in LUAD (Fig. 4), whereas the correlation between *DDP6* and LUAD remains to be further investigated since *DPP6* might have low expression even in the normal lung cells (Fig. S1). The protein expression patterns of *CMTM2* and *TBX4* in LUAD were consistent with their hypermethylation profiles previously identified, confirming that cg08032924 and cg14823851, mapped to *CMTM2* and *TBX4*, are potential novel methylation markers in human LUAD.

### 3.5. *CMTM2* decreases tumor growth of mouse Lewis lung carcinoma *in vivo*

Considering potential clinical significance and biologic implications, we focused on *CMTM2* for further research as the immunohistochemical analysis demonstrates that *CMTM2* is extensively expressed in human lung tissues in our study, implying its comprehensive significance. To elucidate the role of *Cmtm2* in LUAD, we examined the effects of *Cmtm2* on mouse Lewis lung carcinoma (LLC) *in vivo*. The mRNA level of *Cmtm2* revealed that the *Cmtm2* was successfully ectopically expressed in LLC cells (Fig. 5A). Subcutaneously implanted tumor model was applied to C57BL/6J mice to investigate the effects of *Cmtm2 in vivo* (Fig. 5B). Meanwhile, the tumor growth was monitored by measuring the tumor volume (Fig. 5C) as well as living imaging (Fig. 5D,E). It is noticeable that the elevation of *Cmtm2* significantly suppressed the subcutaneous tumor growth compared with the control group (Fig. 5C–F) even in the early stage (Fig. 5D). As expected, the weight of tumors was diminished by *Cmtm2* robustly when the mice were sacrificed after 4 weeks (Fig. 5F,G). These results suggest that *Cmtm2* could suppress the tumor growth of LUAD *in vivo*.

## 4. Discussion

In this study, we explored the global DNA methylation patterns in LUAD and identified three subgroups showing distinct methylation status. CpG island methylator phenotype (CIMP) is characterized by strong hypermethylation of CpG islands in the

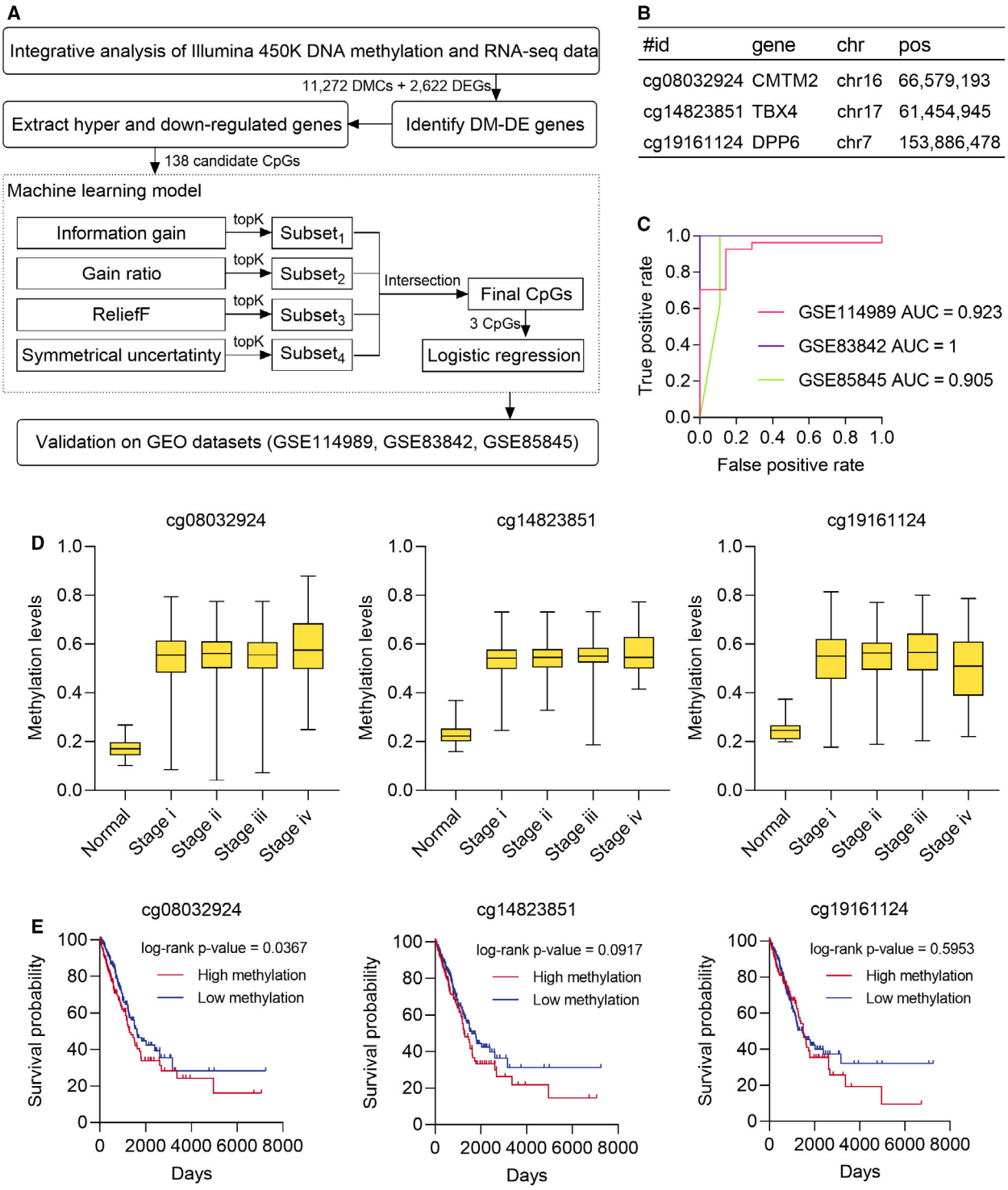


**Fig. 3.** Identification of LUAD methylation markers. (A) The framework of identifying LUAD methylation markers. (B) The genomic details of discovered three methylation markers. (C) Receiver operating characteristic (ROC) curves and AUC values on validation sets. (D) The distribution of methylation levels of three markers in different tumor stages ($n$ = 495). (E) The Kaplan–Meier survival curves for three methylation markers. The boundary for high and low methylation was the average methylation level.

**Table 2.** Top 15 ranked CpGs (genes) identified by different feature selection methods for diagnosis of LUAD.

| IG | GR | SU | RF |
|---|---|---|---|
| ***cg14823851: TBX4*** | ***cg14823851: TBX4*** | ***cg14823851: TBX4*** | ***cg14823851: TBX4*** |
| *cg01158277: CRYAB* | *cg09523275: NKAPL* | *cg10253847: NKAPL* | *cg22620221: DPP6* |
| *cg10253847: NKAPL* | *cg18252309: DPP6* | *cg18252309: DPP6* | ***cg19161124: DPP6*** |
| *cg18252309: DPP6* | *cg10253847: NKAPL* | *cg18694169: NKAPL* | *cg25075794: AQP1* |
| *cg18694169: NKAPL* | *cg18694169: NKAPL* | *cg09523275: NKAPL* | *cg04372674: AQP1* |
| *cg09523275: NKAPL* | ***cg19161124: DPP6*** | ***cg19161124: DPP6*** | *cg01031101: NKAPL* |
| ***cg19161124: DPP6*** | *cg19797376: TAL1* | *cg07153665: CMTM2* | ***cg08032924: CMTM2*** |
| *cg07153665: CMTM2* | *cg07153665: CMTM2* | ***cg08032924: CMTM2*** | *cg18674980: CA3* |
| *cg06499647: C2orf40* | *cg21838979: C2orf40* | *cg19797376: TAL1* | *cg14535980: C2orf40* |
| *cg21838979: C2orf40* | ***cg08032924: CMTM2*** | *cg21838979: C2orf40* | *cg25230363: AQP1* |
| *cg19797376: TAL1* | *cg06499647: C2orf40* | *cg06499647: C2orf40* | *cg10402698: SMAD6* |
| *cg09854734: CMTM2* | *cg05546863: CMTM2* | *cg09854734: CMTM2* | *cg19908768: SULT1C4* |
| ***cg08032924: CMTM2*** | *cg16626067: CMTM2* | *cg05546863: CMTM2* | *cg04567731: TBX4* |
| *cg14535980: C2orf40* | *cg09854734: CMTM2* | *cg14535980: C2orf40* | *cg07510423: C2orf40* |
| *cg17384889: NKAPL* | *cg17384889: NKAPL* | *cg16626067: CMTM2* | *cg01158277: CRYAB* |

The markers commonly identified by four methods (in bold).

promoter regions of tumor suppressor genes, and previous studies have reported that CIMP is associated with patient outcomes in various cancers including colorectal cancer, hepatocellular carcinoma, and gastric cancer [34,35]. The association between CIMP high group and overall survival of LUAD patients remains unclear since there are some discrepancies among different genome-wide methylation studies. Karlsson *et al.* [36] have reported that CIMP shows differences in adenocarcinomas and it is associated with mutation frequency of common tumor suppressor genes such as *KEAP1, TP53, STK11, and SMARCA4*. However, Vaissière *et al.* [33] and Selamat *et al.* [37] have shown that CIMP is unlikely to be present in LUAD. Our analysis results demonstrated that DNA methylation subgroups were associated with genetic and clinical characteristics including sex, stage, smoking history, *KEAP1*, and *STK11* mutation but there is no evidence for poorer overall survival in CIMP group.

Functional annotation analysis revealed that hypermethylated and downregulated genes were enriched in cancer-related pathways such as *Wnt* signaling pathway. A number of researches have shown that *Wnt* signaling pathway is important in the development of lung cancer. Mazieres *et al.* [38] have reported that aberrant methylation of *Wnt* inhibitory factor-1 (*WIF-1*) is an important cause of constitutive activation of the *Wnt* pathway in lung cancer. Selamat *et al.* [37] have shown that sclerostin domain containing 1 (*SOSTDC1*), a secreted regulator of *Wnt* pathway, is hypermethylated and downregulated in LUAD. Consistent with these results, hypermethylation of *WIF-1* and *SOSTDC1* was observed in our study and the mRNA levels of these two genes were decreased in LUAD. Additionally, hypermethylation of other *Wnt* inhibitors including *RSPO1, RSPO2, RSPO4, WNT3A, DKK2, NKD1*, and *TMEM88* was also observed in our study.

To further screen reliable methylation markers, we designed a machine learning framework and identified discriminative CpGs for accurately distinguishing LUAD from normal samples. Feature selection is a machine learning technique, which determines the most relevant features for the target problem [39]. Feature selection has been widely used in biological and medical applications including gene expression analysis [40], transcription factor binding motif analysis [41], and drug discovery [42]. Here, we considered the marker screening as a problem of selecting a relevant feature (CpG) subset for predicting LUAD. Thus, four feature selection models were selected as the initial screening and the results of them were combined together by taking the intersections. Finally, we identified three methylation markers that were cg08032924 (*CMTM2*), cg14823851 (*TBX4*), and cg19161124 (*DPP6*), and the logistic regression model trained using these markers can accurately predict LUAD in independent validation datasets from GEO. *CMTM2* (CKLF Like MARVEL Transmembrane Domain Containing 2) is a protein coding gene that belongs to the chemokine-like factor gene superfamily. *CMTM2* may play an important role in testicular development and is highly expressed in normal adult human testis in a stage-specific manner. Lower expression of *CMTM2* is correlated with spermatogenesis defects including spermatogenesis arrest, which indicates that *CMTM2* might be involved in spermatogenesis [43]. *CMTM2* was found to be expressed at significantly lower level in Sézary syndrome (Sz), an aggressive type of cutaneous T-cell lymphoma, than in benign T-cell samples, and hypermethylation of *CMTM2* promoter can distinguish Sz from erythroderma secondary to

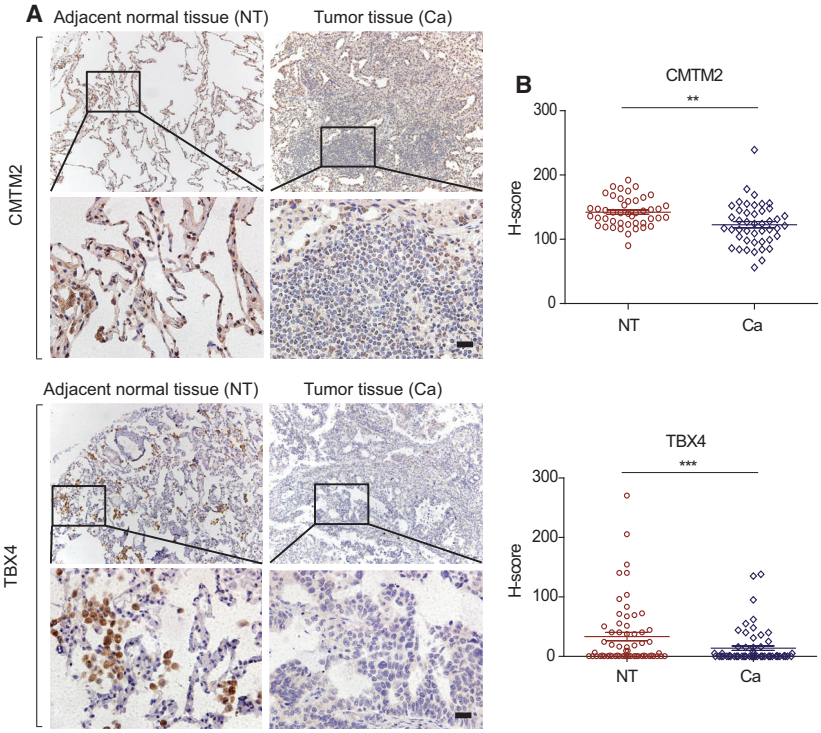


**Fig. 4.** *CMTM2* and *TBX4* are weakly expressed in human LUAD. (A) Representative IHC images of tissue arrays containing human LUAD specimens and paired adjacent normal tissues. Regions in squares are magnified 4× in bottom panels. Scale bar represents 20 μm. (B) Summary statistics of *H*-score based on only intact and paired specimens, $n = 45$ for *CMTM2* and $n = 61$ for *TBX4*. Student's *t*-test. ***P*-value < 0.01, ****P*-value < 0.001.

inflammatory skin diseases [44]. *TBX4* is a transcription factor that belongs to a phylogenetically conserved family of genes that share a common DNA-binding domain, the T-box. *TBX4* was found to be expressed in hindlimb, lung, and proctodeum, and it plays an important role in the development of the hindlimb and in the formation of the umbilicus. A study of a total of 119 bladder cancer samples analyzed by Infinium methylation array showed that *TBX4* was differentially methylated in bladder cancer and was related to disease progression [45]. The methylation of *TBX4* promoter has not been reported to be associated with LUAD, though it has been observed to be downregulated in human non-small-cell lung cancer (NSCLC). Lai *et al.* [46] investigated the expression of a long noncoding RNA *TTTY15* in 37 NSCLC samples and found that downregulation of *TTTY15* was associated with poor prognosis. Interestingly, they also reported that *TTTY15* positively regulated *TBX4* expression by interacting with DNMT3A to affect the binding ability of DNMT3A to the *TBX4* promoter. *DPP6* is a single transmembrane protein that belongs to the peptidase S9B family of serine proteases and is most known for promoting cell surface expression of the potassium channel KCND2. Hypermethylation and decreased expression of *DPP6* were observed in endometrial cancer [47] and melanoma [48] while the role of *DPP6* in LUAD is still unclear.

Early detection, screening, and diagnosis of cancer greatly improve the patient survival rates, as well as significantly reduce the cost and increase the chances for successful treatment. Aberrant DNA methylation plays an important role in cancers and has shown to be a potential biomarker for the early detection of cancer. Compared with other biomarkers such as protein, methylation signature is relatively stable over time and involved in the early stage of carcinogenesis [49]. Moreover, DNA can be isolated with high quality and sufficient yield from frozen biospecimens [50], and DNA methylation status of various gene promoters can be easily captured from biological samples that can be obtained noninvasively including urine, blood, saliva [51]. Epi proColon is an FDA-approved methylation assay that diagnoses colorectal cancer based on methylation status of the target DNA sequence in the promoter region of the *SEPT9*. Methylation markers have also been evaluated for early detection of prostate cancer, and a number of studies have shown that tissue-based *GSTP1* methylation assay can achieve relatively high sensitivity (~ 80%) compared with prostate-specific antigen testing [52,53]. A number of researches have also been done to identify cancer-specific DNA methylation markers in lung cancer patients. Yan *et al.* [54] identified a panel with nine CpGs using a combined public methylation datasets and constructed a prognosis model to predict survival

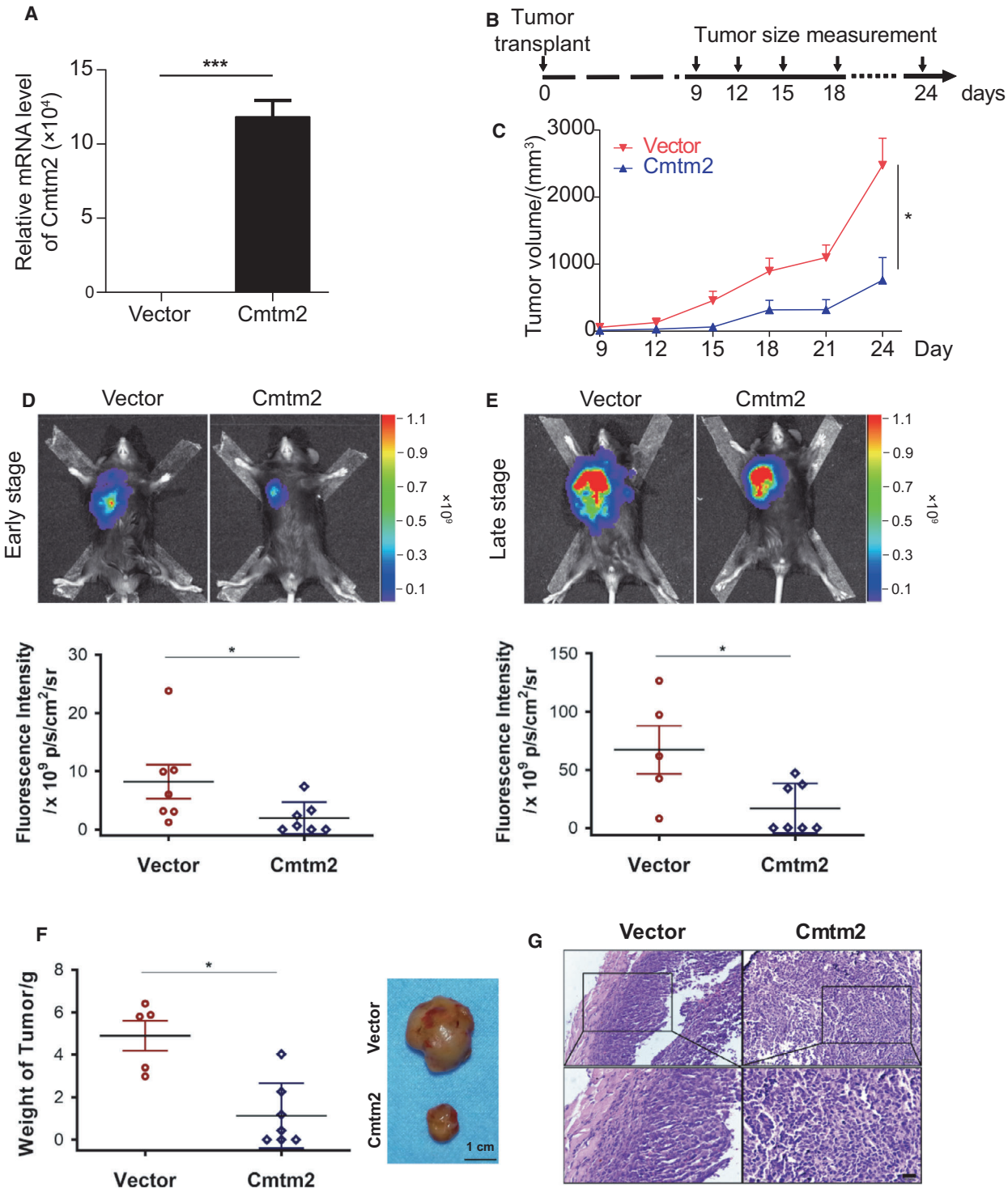


**Fig. 5.** *Cmtm2* decreases tumor growth *in vivo*. (A) The mRNA level of *Cmtm2* is overexpressed in LLC cells ($n = 3$). Student's *t*-test. ***$P$-value < 0.001. (B) Schematic of mouse model on C57BL/6J mice ($n = 14$). (C) Tumor volume of LLC-bearing mice ($n = 14$). Student's *t*-test. *$P$-value < 0.05. (D, E) Representative images and quantification of bioluminescence imaging in early (D) ($n = 14$) and late stage (E) ($n = 12$, 2 died) post-tumor implant. (F) Weight and representative images of excised tumors ($n = 14$). (G) Representative images of HE staining of tumors. Scale bar represents 20 μm. Student's *t*-test. *$P$-value < 0.05.

in LUAD patients. Diaz-Lagares *et al.* [13] identified four methylation markers including *BCAT1*, *CDO1*, *TRIM58*, and *ZNF177* and achieved 85% AUC on a regression model trained from the combination of four markers in bronchoalveolar lavages from patients with lung cancer. In our study, we identified three methylation markers and a logistic regression model trained with these markers on TCGA LUAD samples achieved high AUCs on three independent validation sets. Moreover, we observed these three markers were significantly hypermethylated in stage I LUAD patients, indicating these markers have a great potential to be used to detect LUAD at an early stage. Although the markers have high sensitivity, further validations on different populations are needed and the methylation status of there markers should be validated using cost-effective technology, such as PCR-based methods, in both LUAD tissues and noninvasive samples. In addition, the downstream functions of the marker genes will be systemically studied on our future work.

## 5. Conclusions

In summary, we integrated genome-wide DNA methylation and mRNA expression data and identified three methylation signatures including cg08032924 (*CMTM2*), cg14823851 (*TBX4*), and cg19161124 (*DPP6*) for early diagnosis of LUAD. The results revealed that these markers can distinguish LUAD from normal samples with extremely high AUCs. The decreased expressions of *CMTM2* and *TBX4* were further confirmed in LUAD tissues by IHC. Moreover, we demonstrated that *Cmtm2* could suppress the tumor growth of LUAD *in vivo*. We believe that our study lays the foundation for further biological mechanisms of LUAD development and can contribute to the improvements in early detection and intervention for lung cancer.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (61802209, 81772974), the China Postdoctoral Science Foundation (BS000098), the Fundamental Research Funds of the Central Universities, Nankai University (63181329, 63191422), and the Natural Science Foundation of Tianjin City (16JCYBJC44000).

## Conflict of interest

The authors declare no conflict of interest.

## Author contributions

LW, RX, YS, and YP were responsible for the conception and design of the study. ML and YP contributed to the data analysis. ML, CZ, SL, LZ, and LW conducted the biological validation. ML, YJC, LW, YS, and YP drafted and revised the manuscript. All authors revised, read, and approved the final manuscript.

### Data accessibility

Illumina HumanMethylation450K array, RNA-seq, and clinical information were available at UCSC Xena browser (cohort: GDC TCGA Lung Adenocarcinoma). Three independent validation sets were available at GEO database (accession: GSE114989, GSE83842, and GSE85845).

## Code availability

The source codes are available at: https://sourceforge.net/projects/luad/files/

## References

1 Jemal A, Bray F, Center MM, Ferlay J, Ward E & Forman D (2011) Global cancer statistics. *CA Cancer J Clin* **61**, 69–90.
2 Siegel RL, Miller KD & Jemal A (2020) Cancer statistics, 2020. *CA Cancer J Clin* **70**, 7–30.
3 Molina JR, Yang P, Cassivi SD, Schild SE & Adjei AA (2008) Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clin Proc* **83**, 584–594.
4 Team NLSTR (2011) Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* **365**, 395–409.
5 Black WC, Gareen IF, Soneji SS, Sicks JD, Keeler EB, Aberle DR, Naeim A, Church TR, Silvestri GA & Gorelick J (2014) Cost-effectiveness of CT screening in the National Lung Screening Trial. *N Engl J Med* **371**, 1793–1802.
6 Croswell JM, Baker SG, Marcus PM, Clapp JD & Kramer BS (2010) Cumulative incidence of false-positive test results in lung cancer screening: a randomized trial. *Ann Intern Med* **152**, 505–512.
7 Langevin SM, Kratzke RA & Kelsey KT (2015) Epigenetics of lung cancer. *Transl Res* **165**, 74–90.
8 Ansari J, Shackelford RE & El-Osta H (2016) Epigenetics in non-small cell lung cancer: from basics to therapeutics. *Transl Lung Cancer Res* **5**, 155.
9 Ooki A, Dinalankara W, Marchionni L, Tsay J-CJ, Goparaju C, Maleki Z, Rom WN, Pass HI & Hoque

MO (2018) Epigenetically regulated PAX6 drives cancer cells toward a stem-like state via GLI-SOX2 signaling axis in lung adenocarcinoma. *Oncogene* **37**, 5967–5981.
10 Day JJ & Sweatt JD (2010) DNA methylation and memory formation. *Nat Neurosci* **13**, 1319–1323.
11 Plongthongkum N, Diep DH & Zhang K (2014) Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nat Rev Genet* **15**, 647–661.
12 Laird PW (2003) The power and the promise of DNA methylation markers. *Nat Rev Cancer* **3**, 253–266.
13 Diaz-Lagares A, Mendez-Gonzalez J, Hervas D, Saigi M, Pajares MJ, Garcia D, Crujerias AB, Pio R, Montuenga LM & Zulueta J (2016) A novel epigenetic signature for early diagnosis in lung cancer. *Clin Cancer Res* **22**, 3361–3371.
14 Xu R-h, Wei W, Krawczyk M, Wang W, Luo H, Flagg K, Yi S, Shi W, Quan Q & Li K (2017) Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nat Mater* **16**, 1155–1161.
15 Lee S, Hwang KS, Lee HJ, Kim J-S & Kang GH (2001) Aberrant CpG island hypermethylation of multiple genes in colorectal neoplasia. *Lab Invest* **84**, 884–893.
16 Licchesi JD, Westra WH, Hooker CM & Herman JG (2008) Promoter hypermethylation of hallmark cancer genes in atypical adenomatous hyperplasia of the lung. *Clin Cancer Res* **14**, 2570–2578.
17 Xu W, Xu M, Wang L, Zhou W, Xiang R, Shi Y, Zhang Y & Piao Y (2019) Integrative analysis of DNA methylation and gene expression identified cervical cancer-specific diagnostic biomarkers. *Signal Transduct Target Ther* **4**, 1–11.
18 Goldman M, Craft B, Brooks A, Zhu J & Haussler D (2018) The UCSC Xena Platform for cancer genomics data visualization and interpretation. *BioRxiv* 326470.
19 Hansen KD (2016) IlluminaHumanMethylation450kanno.ilmn12.hg19: annotation for Illumina's 450k methylation arrays. R package version 0.6.0.
20 Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM & Holko M (2012) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* **41**, D991–D995.
21 Dietz S, Lifshitz A, Kazdal D, Harms A, Endris V, Winter H, Stenzinger A, Warth A, Sill M & Tanay A (2019) Global DNA methylation reflects spatial heterogeneity and molecular evolution of lung adenocarcinomas. *Int J Cancer* **144**, 1061–1072.
22 Kajiura K, Masuda K, Naruto T, Kohmoto T, Watabnabe M, Tsuboi M, Takizawa H, Kondo K, Tangoku A & Imoto I (2017) Frequent silencing of the candidate tumor suppressor TRIM58 by promoter methylation in early-stage lung adenocarcinoma. *Oncotarget* **8**, 2890.
23 Yan H, Guan Q, He J, Lin Y, Zhang J, Li H, Liu H, Gu Y, Guo Z & He F (2017) Individualized analysis reveals CpG sites with methylation aberrations in almost all lung adenocarcinoma tissues. *J Transl Med* **15**, 26.
24 Robinson MD, McCarthy DJ & Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140.
25 Network CGAR (2014) Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550.
26 Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC & Lempicki RA (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol* **4**, 1–11.
27 Quinlan JR (1986) Induction of decision trees. *Mach Learn* **1**, 81–106.
28 Quinlan JR (2014) C4. 5: Programs for Machine Learning. Elsevier, Amsterdam.
29 Yu L & Liu H (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. In Proceedings of the 20th International Conference on Machine Learning (ICML-03), pp. 856–863.
30 Robnik-Šikonja M & Kononenko I (2003) Theoretical and empirical analysis of ReliefF and RReliefF. *Mach Learn* **53**, 23–69.
31 Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P & Witten IH (2009) The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsl* **11**, 10–18.
32 Shi Y-X, Wang Y, Li X, Zhang W, Zhou H-H, Yin J-Y & Liu Z-Q (2017) Genome-wide DNA methylation profiling reveals novel epigenetic signatures in squamous cell lung cancer. *BMC Genom* **18**, 901.
33 Vaissière T, Hung RJ, Zaridze D, Moukeria A, Cuenin C, Fasolo V, Ferro G, Paliwal A, Hainaut P & Brennan P (2009) Quantitative analysis of DNA methylation profiles in lung cancer identifies aberrant DNA methylation of specific genes and its association with gender and cancer risk factors. *Cancer Res* **69**, 243–252.
34 Jia M, Gao X, Zhang Y, Hoffmeister M & Brenner H (2016) Different definitions of CpG island methylator phenotype and outcomes of colorectal cancer: a systematic review. *Clin Epigenetics* **8**, 25.
35 Hughes LA, Melotte V, De Schrijver J, De Maat M, Smit VT, Bovée JV, French PJ, Van Den Brandt PA, Schouten LJ & De Meyer T (2013) The CpG island methylator phenotype: what's in a name? *Cancer Res* **73**, 5858–5868.
36 Karlsson A, Jönsson M, Lauss M, Brunnström H, Jönsson P, Borg Å, Jönsson G, Ringnér M, Planck M

& Staaf J (2014) Genome-wide DNA methylation analysis of lung carcinoma reveals one neuroendocrine and four adenocarcinoma epitypes associated with patient outcome. *Clin Cancer Res* **20**, 6127–6140.
37 Selamat SA, Chung BS, Girard L, Zhang W, Zhang Y, Campan M, Siegmund KD, Koss MN, Hagen JA & Lam WL (2012) Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. *Genome Res* **22**, 1197–1211.
38 Mazieres J, He B, You L, Xu Z, Lee AY, Mikami I, Reguart N, Rosell R, McCormick F & Jablons DM (2004) Wnt inhibitory factor-1 is silenced by promoter hypermethylation in human lung cancer. *Cancer Res* **64**, 4717–4720.
39 Piao Y, Piao M & Ryu KH (2017) Multiclass cancer classification using a feature subset-based ensemble from microRNA expression profiles. *Comput Biol Med* **80**, 39–44.
40 Piao Y, Piao M, Park K & Ryu KH (2012) An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. *Bioinformatics* **28**, 3306–3315.
41 Castro-Mondragon JA, Jaeger S, Thieffry D, Thomas-Chollier M & Van Helden J (2017) RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res* **45**, e119.
42 Kang J, Hsu C-H, Wu Q, Liu S, Coster AD, Posner BA, Altschuler SJ & Wu LF (2016) Improving drug discovery with high-content phenotypic screens by systematic selection of reporter cell lines. *Nat Biotechnol* **34**, 70–77.
43 Shi S, Rui M, Han W, Wang Y, Qiu X, Ding P, Zhang P, Zhu X, Zhang Y & Gan Q (2005) CKLFSF2 is highly expressed in testis and can be secreted into the seminiferous tubules. *Int J Biochem Cell Biol* **37**, 1633–1640.
44 van Doorn R, Slieker RC, Boonk SE, Zoutman WH, Goeman JJ, Bagot M, Michel L, Tensen CP, Willemze R & Heijmans BT (2016) Epigenomic analysis of Sezary syndrome defines patterns of aberrant DNA methylation and identifies diagnostic markers. *J Investig Dermatol* **136**, 1876–1884.
45 Reinert T, Modin C, Castano FM, Lamy P, Wojdacz TK, Hansen LL, Wiuf C, Borre M, Dyrskjøt L & Ørntoft TF (2011) Comprehensive genome methylation analysis in bladder cancer: identification and validation of novel methylated genes and application of these as urinary tumor markers. *Clin Cancer Res* **17**, 5582–5592.
46 Lai I-L, Chang Y-S, Chan W-L, Lee Y-T, Yen J-C, Yang C-A, Hung S-Y & Chang J-G (2019) Male-specific long noncoding RNA TTTY15 inhibits non-small cell lung cancer proliferation and metastasis via TBX4. *Int J Mol Sci* **20**, 3473.
47 Huang R-L, Su P-H, Liao Y-P, Wu T-I, Hsu Y-T, Lin W-Y, Wang H-C, Weng Y-C, Ou Y-C & Huang TH-M (2017) Integrated epigenomics analysis reveals a DNA methylation panel for endometrial cancer detection using cervical scrapings. *Clin Cancer Res* **23**, 263–272.
48 Jaeger J, Koczan D, Thiesen H-J, Ibrahim SM, Gross G, Spang R & Kunz M (2007) Gene expression signatures for tumor progression, tumor subtype, and tumor thickness in laser-microdissected melanoma tissues. *Clin Cancer Res* **13**, 806–815.
49 Cheng J, Wei D, Ji Y, Chen L, Yang L, Li G, Wu L, Hou T, Xie L & Ding G (2018) Integrative analysis of DNA methylation and gene expression reveals hepatocellular carcinoma-specific diagnostic biomarkers. *Genome Med* **10**, 42.
50 Jensen SØ, Øgaard N, Ørntoft M-BW, Rasmussen MH, Bramsen JB, Kristensen H, Mouritzen P, Madsen MR, Madsen AH & Sunesen KG (2019) Novel DNA methylation biomarkers show high sensitivity and specificity for blood-based detection of colorectal cancer —a clinical biomarker discovery and validation study. *Clin Epigenetics* **11**, 1–14.
51 Kim H, Wang X & Jin P (2018) Developing DNA methylation-based diagnostic biomarkers. *J Genet Genomics* **45**, 87–97.
52 Van Neste L, Herman JG, Otto G, Bigley JW, Epstein JI & Van Criekinge W (2012) The epigenetic promise for prostate cancer diagnosis. *Prostate* **72**, 1248–1261.
53 Wu T, Giovannucci E, Welge J, Mallick P, Tang W & Ho S (2011) Measurement of GSTP1 promoter methylation in body fluids may complement PSA screening: a meta-analysis. *Brit J Cancer* **105**, 65–73.
54 Yan P, Yang X, Wang J, Wang S & Ren H (2019) A novel CpG island methylation panel predicts survival in lung adenocarcinomas. *Oncol Lett* **18**, 1011–1022.

## Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**Fig. S1.** DPP6 is expressed in a low level.